

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

A Thesis Submitted for the Degree of PhD at the University of Warwick

<http://go.warwick.ac.uk/wrap/55719>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

Applications of Differential Geometry to Statistics.

by

Paul Kenneth Marriott

Thesis submitted for the degree of Doctor of Philosophy
at the University of Warwick

The Mathematics Institute
University of Warwick
Coventry
May 1990.

Summary.

Chapters 1 and 2 are both surveys of the current work in applying geometry to statistics. Chapter 1 is a broad outline of all the work done so far, while Chapter 2 studies, in particular, the work of Amari and that of Lauritzen.

In Chapters 3 and 4 we study some open problems which have been raised by Lauritzen's work. In particular we look in detail at some of the differential geometric theory behind Lauritzen's definition of a Statistical manifold.

The following chapters follow a different line of research. We look at a new non symmetric differential geometric structure which we call a *preferred point manifold*. We show how this structure encompasses the work of Amari and Lauritzen, and how it points the way to many generalizations of their results. In Chapter 5 we define this new structure, and compare it to the Statistical manifold theory. Chapter 6 develops some examples of the new geometry in a statistical context. Chapter 7 starts the development of the pure theory of these preferred point manifolds.

In Chapter 8 we outline possible paths of research in which the new geometry may be applied to statistical theory.

We include, in an appendix, a copy of a joint paper which looks at some direct applications of differential geometry to a statistical problem, in this case it is the problem of the behaviour of the Wald test with nonlinear restriction functions.

Contents.

1. Geometry and Statistics.	
1. Preamble	1.1
2. Basic Statistical objects	1.2
3. Parametric Families	1.4
4. Moments and the Fisher Information	1.8
5. Riemannian Structures	1.9
6. Curvature in Parametric Models	1.15
7. Other Work	1.18
2. Expected Geometry.	
1. Introduction	2.1
2. The Expected Metric	2.1
3. Estimators and the Cramer-Rao Theorem	2.2
4. Limit Theorems and Asymptotic Analysis	2.4
5. The Tangent space and Fisher Metric	2.6
6. Statistical Uses	2.8
7. Non-Metric Connections	2.11
8. The Geometry of α -Connections	2.13
9. Statistical Applications of α -Connections	2.14
10. Lauritzen's Work	2.17
11. Conclusions	2.21
3. Characterisation of Statistical Manifolds	
1. Introduction	3.1
2. Classical Results	3.1
3. Extension to Statistical Manifolds	3.4
4. Conjugate Symmetric Spaces	3.12
5. The Inverse Problem	3.17
6. Conclusion	3.19
4. Equivalences of Statistical Manifolds	
1. Introduction	4.1
2. Statistical Framework	4.1
3. First Order Equivalence	4.2
4. Second Order Equivalence	4.6
5. Third Order Equivalence	4.7
6. Conclusion	4.9

5. Introduction to Preferred Point Geometry	
1. Introduction	5.1
2. Preferred Point Geometry	5.3
3. Power Series Expansions	5.5
4. Examples	5.9
5. Applications to Asymptotic Analysis	5.13
6. Divergence Functions	5.17
7. Choice of Vector Field	5.22
8. Choice of Curvature	5.26
9. Conclusion	5.29
 6. Examples of Preferred Point Geometry	
1. Introduction	6.1
2. The Embedding Metric	6.1
3. Example in Exponential Family	6.5
4. Maximum Likelihood Estimate Geometry	6.7
5. Application to Exponential Family	6.10
6. Expected Log Likelihood Surface	6.11
7. Exponential Family Case	6.14
8. Metrics which produce	
α -connections	6.17
9. Conclusion	6.18
 7. Preferred Point Geometry Theory	
1. Introduction	7.1
2. Comparisons of Riemannian	
and Preferred point Geometries	7.1
3. Curvature in Preferred Point Geometries	7.2
4. Uncertainty	7.4
5. Local Measures of Uncertainty	7.9
6. Uncertainty and Reparametrisation	7.20
7. α -Flat Coordinates	7.22
8. Classification	7.24

8. Future Work.

1. Introduction	8.1
2. Some Questions Raised by this Work	8.2
3. The Role of Geodesics	8.3
4. Maximum Likelihood Estimate Geometry	8.6
5. An Asymptotic Model	8.7
6. Comparison of Estimators	8.13
7. Global Geometry and Singularities	8.14
8. Expected and Observed Geometries	8.16
9. Numerical Work	8.16

Appendix.

References.

Acknowledgments

First, I would like to thank my parents for waiting so patiently. Also I would like to thank my supervisor Prof. J. Eells for letting me try my own way.

A huge amount of thanks must go to Mark Salmon and Frank Critchley without whom there certainly would not be a thesis.

There was much effort put in by my friends proof reading, I would particularly like to thank Duncan and Catharine, Anthony and Maria and Steve for all their work and help. Thanks to Ian and Jane for the chocolate and also to Alison and Guy for keeping me going.

I would like to thank the S.E.R.C. for funding (with Studentship number 85318856).

Chapter One

Geometry and Statistics.

1.1 Preamble.

The aim of this chapter is to describe the motivation and history behind the use of differential geometry in statistics. In writing this work we have a fundamental difficulty in the nature of the audience which might be interested in reading it. On one hand we have the pure geometers. In general the fact that differential geometric theory has found applications in statistics has not been noticed at all in the geometric literature. One reason for this is that very often a geometric audience has a limited grasp of even basic statistical theory, thus the applications are often meaningless to it. Since one of the aims of this work is to introduce the new geometric structures of Statistical Geometry to the wider geometric world we must start by explaining some of the most basic concepts of Statistical Theory.

On the other hand we also have a statistical audience. Here the application and some theory of this geometric approach is much better understood. For example there are many papers describing the area to statisticians who may know little differential geometry, see for example [Amari] or [Lauritzen] and also the collections of papers [Amari 2] and [Dodson]. Because this is the situation at the moment we are going to take the opposite viewpoint from these works. We assume some familiarity with the more geometric definitions and methods, but we shall be much more gentle with any statistical ideas. Therefore we shall make use of elementary statistical examples. We hope any statisticians will just skip quickly through these examples until the more geometric ideas appear. The geometric ideas with which we shall assume familiarity will include the basic notions of tensor analysis as well as calculus of many variables and topology. We shall also assume an understanding of the fundamental concepts of manifold theory.

This introduction will be fairly informal and will not fill in all the details needed. The object being to introduce enough statistical theory to enable any differential geometer to understand the motivation and needs of the statistician. We also review some of the new uses which differential geometry is being applied. For much more detail and a good introduction to general statistical theory we refer to [Cox and Hinkley], [Silvey] and [Hogg and Craig].

Our approach throughout this work is to set up the statistical framework in which we shall be working, hopefully bringing in geometric ideas in a natural way. We try and let the underlying statistical structure determine the geometry. This is particularly true of the new structures that we introduce in chapters 5 and 6. We start by introducing the existing statistical geometric work also in this way, trying to make clear the historical development at the same time. To finish the chapter there will be a description of the rest of the thesis and the motivation for this work.

1.2 Basic Statistical Objects.

We shall be working with some finite dimensional manifolds which naturally occur in statistical theory. Before we start to study them however it is important to understand exactly what a point in one of these manifolds actually is. We shall, therefore, in this section describe some of the basic theory which will enable us to do this.

As a basic reference for this section we shall use and often quote from [Clarke]. Although no doubt almost any standard introductory textbook on probability will contain all the following material.

1.2.1. Definition and Axioms: A *probability space*, (X, \mathfrak{I}, P) is a triple which obeys the following axioms of probability. X is a set which is called the *Sample space* and \mathfrak{I} is a collection of subsets of X called the *Event space*. We have two cases to consider, firstly if X is a discrete set then;

$$\begin{aligned}\mathfrak{I} &= \{ \text{all subsets of } X \} \\ &= \mathcal{P}(X).\end{aligned}$$

If X is a continuous space then \mathfrak{I} satisfies the following;

- (i) $X \in \mathfrak{I}$
- (ii) For any $A \in \mathfrak{I}$ then $X \setminus A \in \mathfrak{I}$
- (iii) For $\{A_i\}$, a countable set of elements of \mathfrak{I} then

$$\bigcup_{i=1}^{\infty} A_i \in \mathfrak{I} \text{ and } \bigcap_{i=1}^{\infty} A_i \in \mathfrak{I}.$$

The third element of the triple P is called a *probability measure* on (X, \mathfrak{S}) and is defined to be a map

$$P: \mathfrak{S} \rightarrow [0,1]$$

which satisfies the following;

- (a) $P(X) = 1$
- (b) If $A, B \in \mathfrak{S}$ and $A \cap B = \emptyset$ then
 $P(A \cup B) = P(A) + P(B)$.
- (c) For $\{A_i\}$ a countable set which is pairwise disjoint
then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Informally we see that the triple (X, \mathfrak{S}, P) consists of a set of events \mathfrak{S} to which the function P is attaching some idea of the probability of occurring. The axioms (i)-(iii) ensure that we can describe the probability of all the events in which we are interested. While the axioms (a), (b) and (c) define the set \mathfrak{S} to be complete in the sense that we are certain that one element of X will happen and that the probability measure is additive in an intuitively nice way. The set X is the set of objects which make up whatever events we are interested in.

We will very often be using the case where X is the real line \mathbb{R} and \mathfrak{S} the set of Borel sets on the real line.

1.2.2 Definition: A real valued function ϕ on X is measurable with respect to the probability space (X, \mathfrak{S}, P) if for all $a \in \mathbb{R}$ we have that,

$$\{x \mid \phi(x) \leq a\} \in \mathfrak{S}$$

1.2.3. Definition: A *random variable* ϕ is a measurable real-valued function

$$\phi: X \rightarrow \mathbb{R}.$$

on (X, \mathfrak{S}, P) , a probability space.

1.2.4. Definition: If ϕ is a random variable on (X, \mathfrak{S}, P) then we define its *Probability Distribution Function* F_{ϕ} by

$$F_{\phi}(a) = P\{\phi \leq a\} = P[\phi^{-1}\{(-\infty, a]\}]$$

1.2.5. Definition: If ϕ is an absolutely continuous random variable then we can define its *probability density function* f_{ϕ} by the following property

$$f_{\phi}(x) = \frac{d}{dx} F_{\phi}(x)$$

where the derivative exists.

1.2.6. Example: One of the most important examples of a probability density function is the *standard normal density*. This is defined to be

$$\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}$$

on the probability space which is given by $\{\mathbf{R}, \mathbf{B}, \mu\}$ where \mathbf{R} is the real line \mathbf{B} the Borel sets and μ is the standard Lebesgue measure.

In general, in this work and in most of the literature, the points which will make up our geometric manifolds will be these density functions. Thus we shall be looking most of the time at finite dimensional function spaces which we can sometimes view as embedded in $C^{\infty}(\mathbf{R}, \mathbf{R})$ or $L^1(\mathbf{R})$ or some other infinite dimensional function space.

1.3 Parametric Families.

We shall now look at the manifolds that we shall be studying in the rest of this work. For an introduction to basic manifold theory for the statistician we refer to the survey paper by Barndorff-Nielsen, Cox and Reid, [B-N,C&R] and also [Kass] or [Murray]. These papers have a particularly statistical viewpoint. For a more general view we shall often refer to the comprehensive [Spivak], alternatively for a good introduction to the differential geometry of surfaces we refer to [do Carmo].

The age and size of the differential geometric literature means

that the style and standard notations which get used in textbooks can vary enormously. We shall mostly work, in the slightly old fashioned way, in explicit coordinate systems rather than the newer, slicker coordinate-free methods. We do this partly because we feel that this makes the material much more accessible to the non-specialist geometer who, we hope, will find this work useful. Also, since this work will have applications in mind, there will often be no choice but to use a concrete coordinate system in which to make our calculations. This approach means that standard theorems which we refer to will often be more easily found in the slightly older texts.

1.3.1. Definition: We have defined a probability density function, $p(x)$, for a random variable. We shall define a *parametric family*, $M=\{p(x,\theta)\}$ to be a set of such densities all with respect to a fixed measure P . They shall be smoothly parametrized by a p -dimensional vector parameter often denoted by $\theta (= (\theta_1, \dots, \theta_p))$.

We see that since we are dealing with a set of densities then they sit naturally in the L^1 space of integrable functions of x . A reference for this and the structure of an L^1 space is [Weir].

1.3.2 Regularity Conditions: All the parametric families, that we shall study will satisfy the following regularity conditions, most of which will ensure that we shall be working on a regular manifold at all times.

- (R1) The domain Θ of the parameter θ is homeomorphic to a p -dimensional Euclidean space \mathbb{R}^p .
- (R2) The topology of the parametric family, M , induced by the parametrisation is compatible with the relative topology of M from its embedding in L^1 .
- (R3) The support of $p(x,\theta)$ is common for all $\theta \in \Theta$, so that $p(x,\theta)$ are mutually absolutely continuous.
- (R4) Every density function $p(x,\theta)$ is a smooth function in θ , uniformly in x , and the partial derivatives $\frac{\partial}{\partial \theta_i}$ and integration of $\ln p(x,\theta)$ with respect to the measure P are always commutative.
- (R5) The moments of the score function $\frac{\partial}{\partial \theta_i} \ln p(x,\theta)$

exist up to the third order and are smooth in θ .

(R6) The Fisher information matrix is positive definite.

In (R6) we refer to the Fisher information matrix. This is one of the most important object in geometrical statistics. We shall define it here but we shall not consider any of its properties until later sections of this chapter and in Chapter 2. The geometric significance of both conditions (R5) and (R6) will become clearer as the chapter progresses.

1.3.3 Definition: The Fisher information matrix is defined to be the matrix given by

$$\left[\int_{\mathcal{X}} \frac{\partial}{\partial \theta_i} \ln p(x, \theta) \frac{\partial}{\partial \theta_j} \ln p(x, \theta) \cdot p(x, \theta) dP \right]_{1 \leq i, j \leq p}$$

We want to show that we have defined parametric families to be regular in such a way that they will form p -dimensional manifolds which lie inside the function space L^1 . To do this we need to show that the map from Θ to the parametrised family is of full rank. This is a consequence of (R6), for details see [Amari].

1.3.4. Example: One the the most important parametric families in statistics is the family of *normal distributions* of which example 1.2.6 is a member. The family is defined by the following set of density functions with respect to the probability space given by $\{\mathbf{R}, \mathcal{B}, \mu\}$;

$$p(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}$$

where $\theta = (\mu, \sigma)$.

This family does fulfill the conditions (R1)-(R6) as can be seen in the reference [Amari].

1.3.5. Example: Another important example is the multinomial family. Here we have $\theta_1, \dots, \theta_{p+1}$ such that

$$\theta_1 + \dots + \theta_{p+1} = 1$$

Then with the sample space being the discrete set $\{1, \dots, p+1\}$ we have the density function

$$p(m, \theta) = \prod_{i=1}^p \theta_i^{\delta(m,i)}$$

where $\delta(m,i)$ is the Dirac delta function.

These examples illustrate the way in which geometry starts to play its part in statistics. Parametric families are natural statistical objects. However the particular type of parametrisation used to name each member of each family is of no statistical importance. Statistical results should often have the property that they are invariant to reparametrisation. It is, however, this property that can be used almost as the definition of geometry. Thus in this fundamental sense we see the basic relationship between geometry and statistics.

Although, as we have stated, the property of invariance is one which we would like to have, in fact, statistical tests do not always. The reason for this is the need for formulas in practical statistics to be tractable. Thus the formulas are frequently approximations and in the approximation invariance can sometimes be lost. One example of this is the Wald test. In the appendix it is shown how a geometric approach can still be used to analyse this lack of invariance and various solutions are proposed to this problem. Thus, even if invariance is not possible to attain, a geometric understanding of the underlying structure can still be very important.

In the regularity conditions (1.3.2) we see that (R1) states that we have parametrised the whole family by a copy of \mathbb{R}^p . We therefore see that, although we have described the result as a manifold, in fact we are dealing with the very simple case of a manifold with a single chart. All of the geometry in this work, and indeed in all of the literature, is of a local nature. It is not clear as yet whether a non-trivial topology can have statistical significance. It is possible to have a manifold with singularities if we relax the condition (R6). It is well known that there is a relationship between identifiability and non-singularity of the Fisher matrix. Therefore it would seem that both a more global type of geometric theory, and the possibility of allowing singularities might be useful extensions of current ideas. However we shall restrict ourselves to a purely local, non-singular analysis of the geometry until the last chapter. In that chapter we shall reconsider this issue when we deal with possible future work.

1.4 Moments and the Fisher Information.

Before we go on with the introduction of the use of geometric structures in statistics it is necessary to have a short review of some statistical objects which we will use later. These objects have a geometric existence and we shall show how Barndorff-Nielsen uses geometric techniques in order to deal with them. A good reference here is [Barndorff-Nielsen and Cox]

1.4.1. Definition: If $p(x, \theta)$ is a density function and $f(x)$ is a random variable then we define the *expected value* of $f(x)$ to be

$$E_{p(x, \theta)}[f(x)] = \int_{\mathcal{X}} f(x) \cdot p(x, \theta) dP$$

We shall often denote this by $E_{\theta}[f(x)]$.

1.4.2. Definition: We define the *first moment* of the random variable, x , which has a density function $p(x, \theta)$, to be the expected value x , i.e.

$$\int_{\mathcal{X}} x \cdot p(x, \theta) dP$$

1.4.3. Definition: We similarly define the *n th moment* of x with respect to $p(x, \theta)$ to be

$$E_{\theta}[x^n].$$

1.4.4. Definition: If we denote the first moment of $p(x, \theta)$ by μ , then the n^{th} central moment of x with respect to $p(x, \theta)$ is defined as

$$E_{\theta}[(x - \mu)^n]$$

We call the first moment and the second central moment the *mean* and *variance* respectively.

These definitions extend to the multivariate case of which we are mostly interested. In particular we shall be interested in the moments of the derivatives of the *log likelihood function*. We shall see this function later in this chapter, for the moment however it is its relationship to the Fisher information and tensor analysis which interests us.

We shall denote the derivatives of the log likelihood function by the following notation;

$$l(x, \theta) = \ln p(x, \theta), \text{ and}$$

$$\partial_i l = \frac{\partial}{\partial \theta_i} \ln p(x, \theta).$$

Barndorff-Nielsen, in [B-N], considers the transformation rules of the moments of the derivatives of the log likelihood function. He finds in this work how the geometric tool of *tensor analysis* can be very useful in the large calculations in which these objects frequently occur. This work can be viewed as seeing how these objects, which are statistical in origin, behave when viewed geometrically. We shall see more of this work in section 1.7.

There is one very important result here, however, which is relevant to our discussion.

1.4.5 Lemma. The Fisher information transforms as a 2-tensor under a change of coordinates (parameters).

Hence the Fisher information is a geometrically well defined object if we wish to use it as a means of measurement. It is in fact a *metric*.

1.5 Riemannian Structures.

We have seen how differential geometry can have applications to the understanding of the problem of invariance in statistics, also how many objects in statistical theory behave like geometric ones. This is however only the beginning of the application of geometry to statistics. What we have seen so far is purely the geometry of differential manifolds. Essentially it is just concerned with how we deal with different ways of parametrising the same geometric objects. There are, of course, many more complex structures in geometry than a manifold. In this section we shall see how *Riemannian Geometry*, that is the geometry of a manifold plus a metric, can be used in statistics.

1.5.1. Definition: A *Riemannian Structure* is a pair (M, g) , where M is a manifold and g is a positive definite symmetric 2-covariant tensor. For details of this structure we refer to [Spivak].

The question of how to give some idea of the distance between points in parametric families, or between any two distributions, is a very natural one in statistics. Attempts have been made to answer this question by the introduction of various *divergence functions* (see [Chentsov]). That is any function which, given two distributions, will give the distance between them. One of the early works in this area is [Jeffreys]. He showed that given a particular divergence function, which we shall call the Jeffreys divergence.

$$\int_x \ln \left\{ \frac{p(x, \theta_1)}{p(x, \theta)} \right\} \{p(x, \theta_1) - p(x, \theta)\} dP$$

then the infinitesimal version is given by the Fisher information. Some other well known divergence functions which have been much studied are the Kullback- Leibler

$$\int_x [\ln p(x, \theta_1) - \ln p(x, \theta)] dP$$

and Hellinger

$$\int_x [\sqrt{p(x, \theta_1)} - \sqrt{p(x, \theta)}] dP$$

These all give the divergence between the two distributions $p(x, \theta)$ and $p(x, \theta_1)$.

It was Rao (see [Rao 1]) who proposed the using the structure of a Riemannian Manifold to extend these ideas of divergence. From a geometric point of view this is the natural structure with which to study distances on parametric families if they are being regarded as manifolds. He proposed the Fisher information matrix as a possible metric and studied some of the Riemannian structures which this produces in examples. Jeffrey's paper [Jeffreys], as we have said, shows the relationship between the Fisher information and his divergence function. In [Rao] the metrics generated by other divergence measures are studied.

In the previous section we have shown that the Fisher information matrix does in fact behave as a metric on a parametric model.

Given any metric on a manifold then we can define the path length of any regular curve with respect to the metric. All Riemannian distances are based on these path integral measures.

1.5.2. Definition: If we have a Riemannian manifold (M, g) and a path on the manifold $\gamma(t)$, i.e. a smooth map

$$\gamma: [0,1] \rightarrow M$$

whose derivative is nowhere zero, then the length of the path is defined to be

$$\int_0^1 \sqrt{g_{ij} \frac{d\gamma^i}{dt} \cdot \frac{d\gamma^j}{dt}} dt$$

1.5.3. Definition: If a and b are two points on a Riemannian manifold (M, g) then the shortest pathlength between them is called the *geodesic distance* between a and b . The path which gives this distance is called the *geodesic* connecting a and b .

1.5.4. Remark: As we remarked earlier on we will almost always be working locally therefore shall not consider here questions of the uniqueness and existence of geodesics. Instead we shall rely heavily on the result that in a small enough open region there always exists a unique geodesic connecting any two points. For more details see [Spivak]

Rao in his early work calculated, in particular examples, formulas for the geodesic distance in the geometry induced by the Fisher metric. Other work in this area includes that by Atkinson and Mitchell [A&M], Burbea and Rao [B&R] also the papers [Skovgaard] and [Eriksen]. However it must be said that apart from particular cases the determination of geodesic distance is not at all an easy problem and in general numerical methods to obtain approximations have to be used.

The Riemannian geometry induced by the Fisher information is one of the most widely studied in the literature. In particular it is the structure studied by Amari. Because the metric is based on the expected information matrix we shall call such a geometry the *expected geometry* of a parametric space. We shall study the results

of using such a structure in Chapter 2. There we shall explain the motivation in using such a metric and also describe the theory and practical applications of the expected system.

This is not, however, the only type of metric which has been studied. Barndorff-Nielsen has produced a geometric system which is based on the observed information matrix. To understand this metric we must look at the concept of *likelihood* and the *maximum likelihood estimator*.

Let $p(x, \theta)$ be a density function. We have been looking at it so far as a real function valued function of θ . We would now like to reverse this viewpoint and treat it with fixed θ as a function of the observed data x . That is we see how $p(x, \theta)$ behaves as we get different values of the random variable x drawn from the sample space. Hence,

1.5.5 Definition: The *likelihood* of $p(x, \theta)$ at a particular x to be

$$\text{lik}(p, x) = p(x, \theta)$$

For reasons that will become clear later it will in fact be more convenient to work with the *log likelihood function* which is the natural logarithm of the likelihood, i.e.

$$l(p, x) = \ln \text{lik}(p, x).$$

For more details on the definition and importance of the likelihood function in statistics we refer to [Cox and Hinkley].

One of the most important problems in statistics is that of *estimation*. Suppose we have a parametric family as our model. That is, we assume that a data generation process lies within our model and is determining the flow of observed data. We also have some fixed observed data from our sample space. The estimation problem is to determine which particular element of the parametric family was the one which governed the data generation process. For details of estimation procedures and theory we refer to [Silvey].

1.5.6. Example: An example of such a problem would be that of a scientist using a theory which predicts the form of probability

distribution which governs the flow of data from his experiment. Let us say that he knows that the data should fit some normal distribution curve. Then the inference problem is to determine which particular normal distribution curve gives the data generation process.

Any procedure for going from a sample to an element of a parametric family is called an *Estimate*. We shall look at a particular estimate as a map from the sample space X to the parametric family M . Perhaps the most important of these estimates is the maximum likelihood estimate.

1.5.7 Definition: Let the observed sample be $x \in X$. Then the *maximum likelihood estimate* (m.l.e.) is the map which takes x to the element of the parametric family M which has the greatest likelihood value for x , i.e.

$$x \rightarrow \{ \hat{\theta}(x) \mid \text{lik}(p(., \hat{\theta}(x)), x) \geq \text{lik}(p(., \theta), x) \text{ for all } \theta \in \Theta \}.$$

Where the m.l.e. for the sample x is $\hat{\theta}(x)$.

Since the natural logarithm is a monotone increasing function the parameter which corresponds to the maximum likelihood function will also be the one which corresponds to the maximum log likelihood. Because of this fact the maximum likelihood estimate is often to be found by solving the *likelihood equations*

$$\frac{\partial l}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \ln p(x, \theta) = 0 \quad (i = 1, \dots, p)$$

We, of course, have to show that the solution to the likelihood equations are a maximum and a global maximum. For regularity conditions and further discussion of the m.l.e. see [Cox and Hinkley] or [Silvey].

We can now see what is meant by the *observed information* which Barndorff-Nielsen uses as a metric.

1.5.8. Definition: The observed information of the m.l.e. $\hat{\theta}(x)$ for the sample x , is the Hessian of the log likelihood function at $\hat{\theta}(x)$, i.e.

$$\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln p(x, \theta) \right]$$

We can view this as a measure of the credence we should give to the maximum likelihood estimate. The larger the information then, informally, the sharper the maximum is. Hence the better the m.l.e. will be in picking out the true data generation process.

Barndorff-Nielsen in [B-N] shows that this is indeed a metric since it transforms under a change of coordinates correctly.

There are major differences between this metric, which we shall call the *observed information metric* and that of the expected information which we saw earlier. The main one is that the observed information metric is dependent on the observed sample x . Because of this it is not clear how to define this metric on the whole manifold M rather than just at the maximum likelihood estimate. Barndorff-Nielsen's solution to this problem is to replace the sample x with the pair (s, a) where s is a sufficient statistic and a an ancillary. For details of this construction see [Cox & Hinkley]. He can then produce a global metric on the parametric family by picking a fixed value for the ancillary statistic.

To end this section we will point out the relationship between the expected and the observed information matrices.

1.5.9 Lemma: The following two expressions are equivalent:

$$E_{\theta} \left[\frac{\partial}{\partial \theta_i} \ln p(x, \theta) \frac{\partial}{\partial \theta_j} \ln p(x, \theta) \right]$$

and

$$-E_{\theta} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln p(x, \theta) \right]$$

Proof: For a proof see [Amari]

Thus we see that the expected information is the expected value of the observed information.

1.6 Curvature in Parametric Models.

Given the existence of a Riemannian structure on a parametric space we have the associated concept of curvature induced by the metric. Curvature is a very involved and complex subject and we shall have to restrict our consideration of it to the simpler issues. We can divide our attention in two ways. The first is the consideration of the curvature of lines and submanifolds within the parametric family itself. This is the so called *covariant curvature*. The other type of curvature we shall consider is the curvature of the manifold itself. We shall look at the *Riemann-Christoffel curvature tensor* to study this. It is here that we miss out a lot of the subtler areas of study in curvature theory, for instance we do not consider in any depth many of the different measures of curvature that exist as contractions of the curvature tensor. The only attempts to do this in the literature are by Lauritzen and Amari who do consider the statistical implications of the *scalar curvature*. We shall look more closely at this in Chapter 2.

We look first at the induced covariant curvature. For details on this subject a good reference is [Dodson & Poston]. To define this curvature we shall in fact define the *connection* which the metric induces. This is the so called Levi-Civita connection. We shall not go into much detail here about the geometric nature of a connection in general, instead we shall refer to the above reference or to [Spivak]. Here are however some basis definitions and properties.

1.6.1. Definition: In a particular coordinate system a connection is defined by its Christoffel symbols Γ_{ij}^k . These obey the following transformation rules for a change of coordinates given by

$$(e_1, \dots, e_p) \rightarrow (\hat{e}_1, \dots, \hat{e}_p),$$

then the Christoffel symbols transform as

$$\hat{\Gamma}_{qr}^s = \sum_{ijk=1}^p \Gamma_{ij}^k \cdot \frac{\partial e_i}{\partial \hat{e}_q} \cdot \frac{\partial e_j}{\partial \hat{e}_r} \cdot \frac{\partial \hat{e}_s}{\partial e_k} + \sum_{\mu=1}^p \frac{\partial^2 e_\mu}{\partial \hat{e}_q \partial \hat{e}_r} \cdot \frac{\partial \hat{e}_s}{\partial e_\mu}$$

where $\hat{\Gamma}_{qr}^s$ are the new symbols.

We define the covariant derivative from the Christoffel symbols by the following formula. If $\gamma(t) = (\gamma_1, \dots, \gamma_p)$ is a curve in (M, g) parametrised by t then in the (e_1, \dots, e_p) coordinate system the covariant derivative is defined by

$$\frac{D}{dt} \gamma^r(t) = \frac{\partial \gamma^i}{\partial e_j} \frac{de_j}{dt} - \Gamma_{rs}^t e_t \frac{de_s}{dt}$$

If we are in a Riemannian manifold there is a particular connection which is called the Levi-Civita connection of the metric which is defined below.

1.6.2. Definition: The Christoffel symbols for the Levi-Civita connection of the metric g are given by;

$$\Gamma_{ij}^k = g^{ks} \Gamma_{ijs} = g^{ks} \cdot \frac{1}{2} \cdot \left(\frac{\partial g_{js}}{\partial \theta_i} + \frac{\partial g_{is}}{\partial \theta_j} - \frac{\partial g_{ij}}{\partial \theta_s} \right)$$

where g^{ij} is the inverse of the metric representative and we are working in the θ -coordinate system.

If a line has zero covariant curvature with respect to a connection then it is said to be a *geodesic with respect to that connection*. The relationship between these geodesics and the ones we have already seen is made clear by the following theorem.

1.6.3. Theorem: If we are in a Riemannian manifold (M, g) then the geodesics with respect to the Levi-Civita connection of g are curves which have minimum length between points.

Proof. See [Spivak]

While the Levi-Civita connection is in many ways geometrically the most natural it is not the only possible connection. There are many other connections on a Riemannian manifold which are not derived in this way from the metric. These connections play a very important part in the use of geometry in statistics.

It was Efron (see [Efron]) who first introduced the concept of a non-metric connection in a Riemannian structure for applications to statistics. He introduced and worked with the family of parametric models which he called *curved exponential models*. The curvature which he defined came about as a measure of information loss in these models. We shall consider this work and the idea of curved exponential families in much greater detail in Chapter 2, where we shall consider it in the context of Amari's framework for non-metric connections in an expected geometry setting.

Independently Chentsov and Dawid introduced a complete 1-parameter family of connections in an expected geometry. For a reference to this we refer to [Chentsov] and [Dawid] and [Dawid 2]. These non-metric connections extend the Efron connection. Again we shall be dealing with these connections and their uses in Chapter 2. Briefly, however, they are the main new tools that geometry has given to statistics. They have been used in the higher order theory of asymptotic expansions, to study the question of how to parametrise a parametric model (see [Hougaard]) and in the study of divergence functions. The theory of such families of connections is new to pure differential geometry theory and so they may have a role to play in further developments there and also perhaps in applications outside statistics.

Barndorff-Nielsen has also developed an observed version of the Chentsov connections.

The other type of curvature which has been considered is the curvature of the whole space or manifold. This use, although clearly connected to the previous ideas of curvature, has had a slightly different history. Curvature in this context has been used to study non-linearity in various models particularly when a theory for a linear model has already been well developed. Bates and Watts (see [B&W]) propose the idea that in a non-linear model care must be taken to differentiate between what they call the *parameter effects curvature* and the *intrinsic curvature* of the model. By this they mean that some

non-linearity can be introduced by the particular form of parametrisation which has been used and this can be eliminated by choosing a better parametrisation. There can also be some non-linearity due to the geometric nature of the manifold itself. They recognised that this can be understood by looking at the Riemann-Christoffel curvature tensor and by trying to find affine coordinates. For more details on this and how the discussion of non-linearity effects an essentially linear test such as the Wald test see appendix.

The existence of a one parameter family of connections also induces a one parameter family of global curvature tensors. These have been used in the study of the reparametrisation problem as we see in Chapter 2. These curvatures have also been used a great deal in the study of the embedding of curved exponential families in full exponential families. Again for more detail see Chapter 2.

1.7 Other Work Involving Geometry in Statistics.

In this section we shall describe other developments in this field which lie a little outside the main body of the literature. We shall not deal with any of this work in the rest of the thesis and so we shall be very brief in the references here.

In his work on observed geometry and on the uses of geometric techniques to handle cumulants, which we have described earlier, Barndorff-Nielsen has also developed a theory of *coordinate strings* these are statistical objects which are generalisations of tensors and other geometric objects. In this work he introduces various operations on these partially geometric objects which he calls *intertwining operations*. These operations have both a statistical and geometric significance. See for more details [B-N]

Murray (see [Murray] and [Murray and Rice]) has developed work which extends the concepts of Taylor series type expansions to a curved manifold structure. Since use of power series methods is very common in statistics this work clearly has many applications.

There have been a number of papers on the uniqueness of the geometric structures which we are introducing. Various reasons as to why the Fisher information is the only plausible (expected) metric and the α -connections are the only non-metric connections, have been put forward. We referred to [Amari] and [Picard]. However, with regard to this work our new structures introduced in Chapter 5 do throw considerable light on these results.

For some collected papers on this subject we can refer to the

two books [Dodson] and [Amari 2] which both contain connected work which we do not describe here.

1.8 Structure of the Thesis.

This thesis can be seen to exist in two distinct, although related, halves. The first half is work which is motivated by the paper [Lauritzen]. In Chapter 2 we look in detail at the expected geometry of Amari and Lauritzen. It is in this framework that Chapters 3 and 4 are set. The work there is involved with the theory of Statistical manifolds (see Chapter 2) and can be seen as answering some questions in the underlying differential geometry which describes the theory. The statistical motivation for this work is set out in Chapter 2 and thus 3 and 4 are mostly pure geometric theory.

The second half of the thesis starts in Chapters 5, 6, and 7. In these chapters we diverge from the current literature and construct what is an essential new geometric structure which we feel is particularly suited to applications to statistics. This geometry, which we call *preferred point geometry*, has the property that it is a non-symmetric structure which reflects the underlying statistical framework. In Chapter 5 we define this structure, using statistical theory to guide the geometry. We show how the new theory is a generalisation of Statistical manifold theory and how it can explain many aspects of the older structure in natural geometric terms. In Chapter 6 we produce some explicit examples of our new geometry and in each case we show the relationship with the expected geometry of Chapter 2, and calculate what the geometry gives us in particular cases, frequently the exponential case. Chapter 7 is concerned with the beginnings of the development of the pure theory of preferred point geometry. We compare aspects of its theory with both Riemannian and Statistical manifold theory and show how the new structure gives useful generalisations of both in a naturally geometric way.

The last chapter of the thesis describes possible future work, in particular new ways of applying our preferred point structure directly to statistics. Also it discusses the ways in which our new geometry does not yet explain all of the previous work and proposes ways in which this might be done. One important area which is discussed is the role of geodesics in statistics and the way in which our new geometry might be able to produce some powerful applications of them.

We include in the appendix the paper [C M & S 1] . This was joint work in which I was involved which analyses the behaviour of the Wald test in the non-linear restriction function case. This work show how geometric analysis can be very useful in direct applications to statistics and it uses the notions of Geodesic inference which we mention in Chapter 8. In this joint work I was responsible for the geometric analysis.

Chapter Two

Expected Geometry.

2.1 Introduction.

In this chapter we are going to review in greater detail than in Chapter 1 the results and motivation behind the theory of expected geometry. The first half of the chapter will be a review of the work of Amari. Many of the proofs and much more detail will be found in his book [Amari] and [Kumon & Amari]. We follow his work in detail here since the results in Chapters 5, 6 and 7 can be seen as generalisations of Amari's theory of Statistical geometry and his work provided the motivation for ours.

The second half of this chapter is a review of the work on statistical geometry by Lauritzen. This follows very naturally from the work of Amari and seeks to provide a framework which will contain both the expected and the observed theories. The reason for looking at this work in detail is that at the end of the paper [Lauritzen], Lauritzen poses some open problems, two of which we solve in Chapters 3 and 4.

2.2 The Expected Metric.

We have seen in Chapter 1 how the Fisher information can be seen as a metric on a manifold which is given by a parametric family. Before we start to study the results of this expected geometry we shall look at the statistical motivation for using the Fisher information metric.

We have seen that one viewpoint is to note that the Fisher information is the local version of some already existing divergence function. This is the result of the papers by Rao and by Jeffreys.

We shall look at two other ways of seeing the statistical behaviour of the Fisher information. The first based on the behaviour of estimates and the Cramer-Rao theorem. The second based on the asymptotic performance of various estimators. The above three ways of viewing the Fisher information are reflected in the applications of the expected geometry which it generates.

2.3 Estimators and the Cramer-Rao Theorem.

We have already seen one example of an estimator in Chapter 1. That was the maximum likelihood estimator. This has become the most important estimator in statistics. To understand why this is so and how to compare the performance of different estimators we need to study the Cramer-Rao theorem. A good reference for this topic is the book [Silvey].

We shall first give an example of another commonly used estimator for comparison.

2.3.1 Definition: The method of moments is a method of estimating parameters based on equating population and sample values of certain moments of a distribution. For example, in the univariate case the h parameters $\theta_1, \dots, \theta_h$ of the density $f(x; \theta)$ have moment estimates given by solving

$$\mu'_r(\theta) = m'_r \quad r = 1, \dots, k$$

where m'_r is the r^{th} noncentral sample moment and

$$\mu'_r(\theta) = \int_a^b x^r f(x, \theta) dx \quad r = 1, \dots, k \quad k \geq h$$

is assumed to exist. If $k > h$ then the values of r used above need not be consecutive. For more details see [Kotz and Johnson].

Given that there are different estimators used in statistics the question arises as to which is the 'best'? To make sense of this we have to clarify what we mean by such a question. Once answered we can use the solution to give a form of distance measurement on our manifold.

For the moment to avoid pathological estimators which have no statistical use we shall restrict the possible estimators to an important subclass. We shall look at the class of *unbiased* estimators. Where we define an estimator to be unbiased if the expected value of the estimator, with respect to a particular distribution, is the parameter of the distribution itself. This is summed up in the formula

$$\forall \theta \in \Theta, \quad E_\theta[\bar{\theta}(x)] = \theta$$

where $\bar{\theta}(x) \in \Theta$ is our estimator for the data x .

In this subspace of unbiased estimators we have a statistical theorem which tell us how good an estimator can be. This is the Cramer-Rao theorem. We shall state it and then describe what it means in this context.

2.3.2 Theorem (Cramer-Rao): Let $\hat{\theta}$ be an unbiased estimate and θ the true parameter. Denote the variance-covariance matrix of $\hat{\theta}$ by $\text{var}_{\theta}(\hat{\theta})$. Then the *Fisher Information*, which is defined to be the matrix,

$$I_{\theta} = E_{\theta} \left[\frac{\partial}{\partial \theta_i} \ln p(x, \theta) \frac{\partial}{\partial \theta_j} \ln p(x, \theta) \right]$$

has the property that $(\text{var}_{\theta}(\hat{\theta}) - I_{\theta}^{-1})$ is a positive semi-definite matrix.

We should view the above result as giving a bound to the variance of any unbiased estimator. The variance measures the 'spread' of a distribution and we want estimators to have as small a variance as possible. Hence the theorem gives us a bound for the variance of the 'best possible' choice of estimate.

The question remaining is, when can this bound be obtained? It is not true, in general, that it can always be reached. However as the sample size tends to infinity we do know that the maximum likelihood estimate tends to this bound. For more details of this see [Silvey].

We can now get an intuitive feel for the use of the Fisher information as a basis for making distance measurements on our parametric family. Let us suppose that we have an inference problem on our family and that the true parameter θ_0 represents the data generation process. Now the Fisher information at θ_0 will give a measure of how well the best possible estimator will be able to distinguish between θ_0 and any nearby point. Thus the 'distance' that the Fisher information is giving is a measure of how well we can distinguish between the true parameter and any other if we are using the best possible method of separating the two distributions.

We should view this distance as a theoretical best case, which may or may not be relevant to a particular statistical problem. One reason for this is because we are dealing with an expected measure. Thus we are thinking all the time about an expected type of distance

measure. Areas where this measure may not be relevant are the so called *finite sample* problems and also when the variance of the actual best possible estimate is bounded away from the Cramer-Rao lower bound (see [Silvey]).

As long as we are aware of its restrictions we still find that the Fisher information is an extremely useful tool for the statistician.

2.4 Limit Theorems and Asymptotic Analysis.

We shall in this section look at another purely statistical use of the Fisher information which will give us the motivation for using it as a metric. In order to do this and to understand much of the work of Amari we look at some statistical theory to which the geometric approach can give very useful insights.

The first topic we shall look at is *repeated sampling*. We shall have an underlying probability space (X, \mathfrak{F}, P) where the associated density function $p(x, \theta_0)$ lies in a regular parametric family M . We shall call θ_0 the true parameter. Suppose now that rather than just take one element of our sample space x , we repeatedly take samples and get a sequence $\{x_i\}$. We shall denote the (finite) sequence $\{x_i\}$ by the n -vector \underline{x} (or \underline{x}_n) and we shall let $\underline{x}_n \in X^n$, say. We now want to find the function determining the probability density of \underline{x} . Since each element of \underline{x} has the same density function $p(x, \theta_0)$ the sequence is called *independently identically distributed* (i.i.d.). For this situation the density function of \underline{x} is easy to calculate as

$$p(\underline{x}_n, \theta_0) = \prod_{i=1}^n p(x_i, \theta_0)$$

If we work with log-likelihoods, as in Chapter 1, then the formula is even simpler

$$l(\underline{x}_n, \theta_0) = \ln p(\underline{x}_n, \theta_0) = \sum_{i=1}^n \ln p(x_i, \theta_0) = \sum_{i=1}^n l(x_i, \theta_0)$$

This is, in fact, one of the reasons for using log-likelihoods rather than just the likelihood function.

2.4.1. Example: Recall example 1.5.6 about estimation. We would

have i.i.d. sampling if the scientist takes repeated readings from his experiment and then works with the assumption that the distribution which governs the variability in these readings does not vary with time.

We shall now look at the most basic behaviour of sequences of random variables. In particular here and, in fact, for the rest of the thesis we shall work with the assumption that we have i.i.d. sampling. This is a very common assumption in all the literature on the expected geometry in statistics.

We shall first look at the types of convergence in statistics. For more details about this and indeed this whole section see [Prakasa Rao]

2.4.2. Definition: A sequence of random variables $\{\varphi_n\}$ which have distribution functions F_n is said to converge in law (distribution) to φ with distribution function F if

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

at all continuity points x of F .

We denote this convergence by

$$\varphi_n \xrightarrow{\text{law}} \varphi.$$

and F_n is said to converge weakly to F ($F_n \xrightarrow{w} F$).

As we increase the sample size in our i.i.d. distribution of \underline{x} it is a very important issue in statistics to see how the distribution converges. We shall now quote the most important of all the results in this area which is called the *Central Limit Theorem*.

2.4.3. Theorem:(The Central Limit Theorem) Let $\{\phi_n \mid n \geq 1\}$ be i.i.d. random variables with mean μ and variance σ^2 . Let

$$S_n = \phi_1 + \phi_2 + \dots + \phi_n,$$

then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{\text{law}} N(0, I).$$

Where $N(0, I)$ is the standard normal (multivariate) distribution defined in 1.2.6.

This limit theorem is the basis of what is known as asymptotic analysis. The idea is to get asymptotic expansions of the distribution of objects like S_n in terms of powers of $(1/n)^{1/2}$, where n is known as the *sample size*. These expansions enable us to get a better understanding of the distributions of sums of i.i.d. random variables by taking higher order terms. The limit theorems we have just seen can be understood as giving us the first (and limiting order) terms of these expansions.

The most important of these expansions is known as an Edgeworth expansion.

2.4.4. Definition: Using the notation of 2.4.3. we have the following asymptotic expansion of F_n , the distribution function of the normalised sum of i.i.d. random variables with mean μ and variance σ^2 .

$$\begin{aligned} F_n(x) &= \Phi(x) - \sum_{j=1}^{\infty} \frac{P_j(\Phi)}{n^{j/2}} \\ &= \Phi(x) - \frac{\lambda_3 \Phi^{(3)}(x)}{6\sqrt{n}} + \frac{1}{n} \left(\frac{\lambda_4 \Phi^{(4)}(x)}{24} + \frac{\lambda_3^2 \Phi^{(6)}(x)}{72} + \dots \right) + \dots \end{aligned}$$

Here $\Phi(x)$ is the standard normal distribution, $\Phi^{(k)}$ denotes the k th derivative of Φ and λ_j denotes the j th order *cumulant* of $p(x, \theta_0)$ the underlying distribution. P_j is a polynomial of degree $3j$ whose coefficients depends on the cumulants λ .

For more of the exact details which do not concern us here see [Prakasa Rao] or [Amari].

2.5 Representations of the Tangent Space and the Fisher Metric.

We can now return to the geometric development and look at the third of our ways of viewing the Fisher information as a metric. We shall look here at the tangent space to our manifold and see how the Fisher information acts on it.

As with the likelihood we shall find that working with the logarithms has many advantages. Therefore, instead of considering the manifold given by the family

$$\{ p(x, \theta) \} \quad (N)$$

instead we consider

$$\{ \ln p(x, \theta) \}. \quad (1)$$

There is clearly an isomorphism between the corresponding tangent spaces. We shall deal with the tangent space to (1). We follow Amari in that we call this the *1-representation of the tangent space*.

We can see that there is a direct use of the Fisher information as a quadratic form on the 1-representation of the tangent space. This comes from the asymptotic behaviour as the sample size tends to infinity.

The 1-representation of the tangent space means we are dealing with the linear space at θ generated by the functions;

$$\left\{ \frac{\partial}{\partial \theta_i} \ln p(x, \theta) \right\}_{i=1 \text{ to } p}$$

where we are working in the θ -coordinate system.

Rather than just working with the single observed sample we are working in the i.i.d. repeated sample. Thus we are looking at the manifold still parametrised by θ except with the densities $\{p(\underline{x}_n, \theta)\}$. This is, as we have said in section 2.5, a very common occurrence and for the rest of this work this is the situation that we are interested in. The 1-representation of the tangent space of this new manifold has a very simple relationship to the original since,

$$l(\underline{x}_n, \theta_0) = \ln p(\underline{x}_n, \theta_0) = \sum_{i=1}^n \ln p(x_i, \theta_0) = \sum_{i=1}^n l(x_i, \theta_0)$$

hence,

$$\frac{\partial}{\partial \theta_i} \ln p(\underline{x}_n, \theta_0) = \sum_{k=1}^n \frac{\partial}{\partial \theta_i} l(x_k, \theta_0)$$

Therefore an element of the tangent space is a sum of i.i.d. random variables. Hence we can apply the central limit theorem of Section 2.4. In the asymptotic limit the vectors in the tangent space at θ_0 have a normal distribution with variance which when calculated is the Fisher information.

This gives us the third way of viewing the Fisher information as a metric on a manifold. The variance of the distribution of the tangent vectors is frequently taken as a measure of the spread of the distribution in statistics. Hence it is natural to use it as a measure on the tangent space at the true parameter.

To sum up the previous four sections, we can see that the Fisher information or expected geometry can be seen in various lights. Firstly, following Rao, as an infinitesimal version of already existing divergence functions. Secondly, as a geometry which is determined by the (expected) indistinguishability of densities. When in this case we are doing estimation, the Cramer-Rao theorem gives us the motivation. Thirdly the viewpoint is an asymptotic one where we are viewing the behaviour of elements of the tangent space

2.6 Statistical uses of the Fisher Information metric.

After this general discussion of the statistical role of the Fisher information and its geometric significance we shall continue with the applications of expected geometry. We shall be following the work of Amari in the next few section and for more details we refer to the book [Amari].

To start with we shall introduce the classes of parametric families, where most of Amari's work has its natural setting. These classes are known as *full and curved exponential families*. While the class of full exponential families has been well known and important in classical statistics, it was Efron (see [Efron]) who first set out the generalisation to the curved case and started the analysis of it in geometric terms.

2.6.1. Definition: The *exponential families* are characterised by having probability density functions of the form:

$$p(x, \omega) = a(\omega)b(x)\exp\{\theta(\omega).t(x)\}$$

where ω is a parameter and $\theta(\omega)$ and $t(x)$ are vectors of common dimension, k say, and $.$ denotes inner product. Let P denote the exponential family of distributions with density functions $p(x, \theta)$. The probability functions are all densities with respect to the same measure μ , which is typically a Lebesgue measure. Let Ω be the domain of variation for θ and let $\Theta = \theta(\Omega)$ denote the canonical parameter domain for P . Let

$$\hat{\Theta} = \left\{ \theta: \int b(x)\exp(\theta.t) d\mu < \infty \right\}$$

which is a convex subset of \mathbb{R}^k . Then P is said to be *full* if $\Theta = \hat{\Theta}$ and P is *regular* if it is full and an open subset of \mathbb{R}^k . For more details see [Kotz and Johnson]

2.6.2. Definition: If Θ is a smooth submanifold in $\hat{\Theta}$ then it said to be a *curved exponential family*. We denote it by the set of densities

$$p(x, \theta) = \exp\left\{ \sum_{i=1}^m u_i(\theta)x^i + b(x) - \psi(\theta) \right\} \quad (i = 1 \text{ to } m)$$

where $u(\theta)$ is an m -dimensional function of the p -dimensional parameter θ and $p \leq m$. The sample space X is m -dimensional and $x \in X$ has the form $x = (x_1, \dots, x_m)$.

The function $\theta \rightarrow u(\theta)$ defines an immersion of the curved exponential family into a larger dimensional full exponential family. Thus a curved exponential family is (modulo regularity conditions) a submanifold of a full one.

Examples of the families include the normal and multivariate normal which we have already seen.

Since the Fisher information induces a Riemannian structure we can do the usual calculations on any of our parametric families. For example we can calculate the induced Levi-Civita connection, and the Riemann-Christoffel curvature tensor for the manifold.

Amari has done this for many cases. One particularly interesting result, from a geometric point of view, is the fact that a simple

calculation shows that the family of normal distributions which we defined in 2.1.4 is a manifold of constant negative curvature.

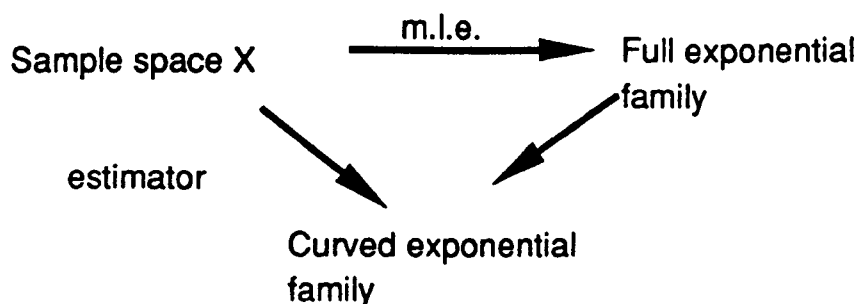
Since the normal family is a hyperbolic space its geometry is well known. In particular its geodesics, and the corresponding geodesic distances, have been calculated. This is not true of many other families although the geodesics for the multivariate case have been studied in [Dodson].

Another use for the induced curvature tensor for the Fisher geometry has been its role in seeing how good a parametrisation can be achieved. This was studied in [Jeffreys] where a zero curvature tensor was used to indicate the existence of a *covariance stabilising* coordinate system.

Apart from these results on the curvature of the Fisher geometry the metric itself has its uses in the calculation of the efficiency of estimators in curved or full exponential families. This is the approach to geometry put forward mostly by Amari, see [Amari]. We shall briefly describe the basic results here.

The term efficiency refers to the asymptotic efficiency of an estimator on a curved exponential family.

Geometrically a curved exponential family is embedded in a full exponential family. We look at the case where we are doing estimation on the curved submanifold. The method which Amari analyses is to use the m.l.e. in the full exponential family in all cases. All estimation maps from the sample space to the curved exponential submanifold are thought of as factoring through the m.l.e. estimator to the full family. In other words the estimator must belong to the following commutative diagram;

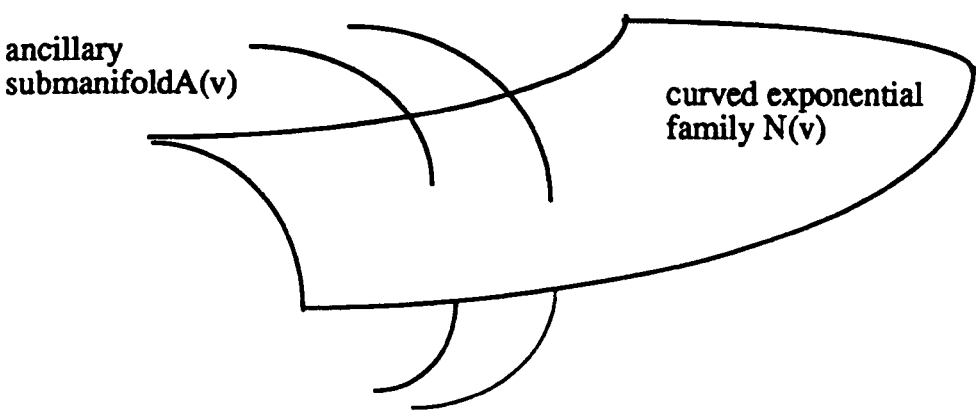


Let the full exponential family be M and the curved submanifold be N . The commutative diagram above defines a map $\phi: M \rightarrow N$ for

each estimator $\Phi:X \rightarrow N$. Φ maps a manifold onto a submanifold hence to each point v on the submanifold there exists an ancillary submanifold consisting of all the points mapped to v (see figure below). That is the fibre of Φ which is

$$A(v)=\{\Phi^{-1}(v)| v\in N\}.$$

The following diagram shows a two dimensional curved exponential family embedded in a three dimensional full exponential family thus the fibres $A(v)$ are one dimensional.



Amari makes use of the correspondence between the ancillary submanifolds through N and the estimators of N . He proves results on the behaviour of the estimators of N by proving geometric results on the corresponding ancillary families.

We shall now state Amari's results concerning Fisher information and estimators.

2.6.3 Theorem: An estimator is consistent, when and only when for every point, $v\in N$, the submanifold, is included in the associated ancillary submanifold $A(v)$ attached to v .

2.6.4. Definition: Let the mean square error of a consistent estimator have the following power series expansion in $n^{-1/2}$

$$g_1^{ab}(v) + g_2^{ab}(v) \frac{1}{\sqrt{n}} + g_3^{ab}(v) \frac{1}{n} + O(n^{-\frac{3}{2}})$$

The estimator is said to be *first order efficient* when it has a minimal first order term among all other consistent estimators.

Again for details see [Amari].

2.6.5. Theorem: A consistent estimator on N is first order efficient if and only if its associated ancillary family cuts the submanifold orthogonally in the Fisher metric.

Thus we see that it is the metric on the submanifold which determines the statistical efficiency of an estimator.

2.7 Non-metric Connections.

We have introduced a way of producing a Riemannian structure on a statistical space. However it has been one of the major results of the geometrisation of statistics that this structure alone wasn't enough to contain all the statistical information in a parametric space. It was realised that a Riemannian structure was not enough and the geometric concept of a connection had to be added if we wished to extend and generalise the previous results. The connections which are added are not the induced metric connection for the Fisher metric but are related to the metric in an interesting way.

Following Amari we shall introduce the idea of a *one-parameter family of connections on a manifold*. These connections can be parametrised by α which is a real number. Thus we talk about the α -connection on the manifold. All proofs in this section are straightforward calculations and can be found in [Amari].

2.7.1 Notation: In the θ -coordinates we denote the Christoffel symbols for the Levi-Civita connection for the Fisher metric by;

$$\Gamma_{ijk}(\theta) \quad (1 \leq i, j, k \leq p)$$

2.7.2 Definition: We define the *skewness tensor* T by the following,

$$T_{ijk}(\theta) = \int_{\mathbf{x}} \frac{\partial \ln p(\mathbf{x}, \theta)}{\partial \theta_i} \cdot \frac{\partial \ln p(\mathbf{x}, \theta)}{\partial \theta_j} \cdot \frac{\partial \ln p(\mathbf{x}, \theta)}{\partial \theta_k} \cdot p(\mathbf{x}, \theta) d\mathbf{x}$$

where we are integrating over the sample space. Alternatively this can be written as

$$T_{ijk}(\theta) = E_{\theta} \left[\frac{\partial \ln p(\mathbf{x}, \theta)}{\partial \theta_i} \frac{\partial \ln p(\mathbf{x}, \theta)}{\partial \theta_j} \cdot \frac{\partial \ln p(\mathbf{x}, \theta)}{\partial \theta_k} \right].$$

2.7.3 Lemma: T is a symmetric 3-tensor.

We can now define the α -connection by its Christoffel symbols in this coordinate system.

2.7.4 Definition: For $\alpha \in \mathbb{R}$ we define the Christoffel symbols of the α -connection by

$$\Gamma_{ijk}^{\alpha} = \Gamma_{ijk} - \frac{\alpha}{2} T_{ijk}(\theta)$$

2.7.5 Lemma: The above definition does define a set of Christoffel symbols for a well-defined connection i.e., the definition does not depend on the coordinate system chosen.

The one parameter family of connections enables us to define a corresponding family of geometric objects. Thus for each α we have a set of corresponding geodesics and also we can define the curvature tensor for each connection. In the obvious way these are called the α -geodesics and α -connections.

2.7.6. Definition: We recall the curvature tensor with respect to the α -connection is defined to be

$$R_{ijkm}^{\alpha} = \left(\partial_i \Gamma_{jk}^{\alpha s} - \partial_j \Gamma_{ik}^{\alpha s} \right) g_{sm} + \left(\Gamma_{irm}^{\alpha} \Gamma_{jk}^{\alpha r} - \Gamma_{jrm}^{\alpha} \Gamma_{ik}^{\alpha r} \right)$$

We see that although none (except for $\alpha=0$) of these connections are metric, however the Fisher metric is tied up closely in

their definition. We see, first, that the 0-connection is the Levi-Civita connection for the Fisher metric. We shall now state more of the purely geometric results on the family of connections

2.8 The Geometry of the Family of α -Connections.

The family of α -connection has a rich internal structure which is new to pure differential geometry theory. We shall in this section state the most important results concerning the relationship between the non-metric connection and the metric and also the new topic of *duality* in such a family.

2.8.1. Definition: For each $\alpha (\in \mathbb{R})$ the $-\alpha$ -connection is said to be the dual connection to the α -connection.

2.8.2. Note: The 0-connection, which is the Levi-Civita connection for the Fisher metric, is therefore the only self-dual connection in the family.

We denote the metric at a point θ by the inner product

$$\langle \cdot, \cdot \rangle_{\theta}: TM_{\theta} \times TM_{\theta} \rightarrow \mathbb{R}.$$

where TM_{θ} denotes the tangent space to the manifold at θ . Further, we denote the parallel displacement with respect to the α -connection of a tangent vector, v , along a path γ by

$$\Pi_{\alpha}(\gamma, v).$$

Then we have that the α -connections obey the following form of duality.

2.8.3. Theorem: For all $\alpha \in [-1, 1]$ and for all paths $\gamma(t)$ which start at θ , if v and w are two tangent vectors in the same tangent space T_{θ} then,

$$\langle v, w \rangle_{\theta} = \langle \Pi_{\alpha}(\gamma(t), v), \Pi_{-\alpha}(\gamma(t), w) \rangle_{\gamma(t)}$$

Proof. See [Amari] Chapter 3.

For each α we have a geometric structure. In particular we have a curvature tensor T^α which is induced from the α -curvature in the natural way. The duality relationship between the α and $-\alpha$ case extends to the α -curvature in the following way;

2.8.4 Theorem: If the α -curvature of a parametric family is zero then the $-\alpha$ -curvature of the family is also zero.

Proof. [Amari] Chapter 3.

2.9 Statistical Applications of α -Connections.

We have seen some of the basic geometry of the α -connection. We shall now list some of the applications which these connections have been put to in statistical theory, again leaving out most of the proofs and giving references where they can be easily found.

The first application we shall look at is Amari's use in the higher order theory of asymptotic expansions. We have already seen in 2.6.1 how the Fisher metric has been used to calculate the first order efficiency of an estimator for a curved exponential family embedded in a full exponential family. Amari shows to extend this theory to the higher order terms of the asymptotic expansion. The pure Riemannian structure is not sufficient and it is necessary to introduce the α -connections to extend the geometric theory. We shall state here his results.

We work in the same framework as in 2.6.3 and 2.6.5 and look at the higher order terms of that expansion in a geometric way.

2.9.1. Definition: If a first order efficient estimator has a second order term in the expansion of 2.6.4 which is minimal among all first order efficient estimators, then the estimator is said to be *second order efficient*. If the third order term is minimal among all second order efficient estimators then the estimator is said to be *third order efficient*.

2.9.2. Theorem: A bias corrected first order efficient estimator is always second order efficient.

2.9.3 Theorem: A second order efficient estimator is third order efficient if and only if the -1 -curvature of its associated ancillary family

is zero.

Thus here we can see one use of the α -connections. Their use has, however, not been restricted to the analysis of asymptotic expansions. We shall see how the induced α -curvature can be used as a measure of whether or not certain reparametrisations are available for parametric families.

If, for some value of α , the α -curvature tensor is zero then, by a standard theorem of differential geometry (see [Spivak]) we know there exists a set of coordinates which are affine with respect to this connection. Amari shows how these affine coordinates have particularly nice properties when they are used to parametrise a family of distributions. We list here the results for the values of α that Amari considered.

2.9.4. Theorem: Characteristic features of α -affine parameters for an exponential family are

- 1) When $\alpha=1$ it is a (locally) natural parameter.
- 2) When $\alpha=1/3$, it is a (locally) normal likelihood parameter
- 3) When $\alpha=0$, it is a (locally) covariance stabilizing parameter.
- 4) When $\alpha=-1/3$, it is a (locally) zero skewness parameter.
- 5) When $\alpha=-1$, it is a locally minimum covariance parameter.

The role of the duality, which is implicit in the geometry of the family of α -connections, has an interesting parallel in applications for particular statistical families. In particular we note the following facts;

2.9.5 Theorem: The natural parameters for an exponential family (see [Amari]) are affine parameters for the connection induced for $\alpha=1$.

2.9.6 Theorem: The natural parameters for a mixture family, i.e., the mixing parameters, are affine for the connection induced for $\alpha=-1$.

Thus we see that, in this sense, the exponential and mixture families, which are two of the most important sets of families of

distributions in statistics, are dual.

We also have, directly from the above theorems and Theorem 2.8.4, the following result.

2.9.7 Theorem: Any exponential family and any family which is parametrised by linear mixture coefficients, is flat in the geometry induced by the 1 and the -1-connections.

We also have another interesting result along these lines about the family of normal distributions.

2.9.8 Theorem: The 2-manifold which is given by the family of normal distributions $N(\mu, \sigma)$ has constant scalar curvature for the α -curvature tensor for all values of α .

This is another result for which it would be interesting to have a statistical interpretation.

The other main use of the α -connections, in Amari's work, has been in the study of divergence functions. He shows that in an α -flat family there exists a divergence function which is consistent with the α -structure in the following sense. Let M be an α -flat family and N a submanifold. For a point $\theta \in M$ sufficiently close to N there is a unique projection from θ to N using the α -geodesic from θ which cuts N orthogonally. Amari shows that this projection minimises the divergence function evaluated at θ and any point of N . He goes on to show that for a -1-flat family, say an exponential one, the divergence function which corresponds to this connection is the Kullback-Leibler divergence.

2.10 Lauritzen's Work.

We shall now look at the work of Lauritzen in the field of statistical geometry. While his work does encompass the expected geometry that we have been looking at here it also provides a unifying framework to include both expected and observed geometry. We include it here partly because it proposes a new theory which is directly applicable to the expected geometry theory. Also, however, in [Lauritzen] there are a number of open questions about this framework which we tackle in Chapters 3 and 4.

2.10.1. Definition: We define a *Statistical manifold* to be a Riemannian manifold with a symmetric covariant 3-tensor D . We write it as (M, g, D) where M is the manifold and g is the metric. We call D the *skewness* of the manifold.

Thus we see immediately that Amari's structure fits into this framework where the metric is the Fisher metric and D is given by the tensor

$$T_{ijk}(\theta) = E_{\theta} \left[\frac{\partial \ln p(x, \theta)}{\partial \theta_i} \frac{\partial \ln p(x, \theta)}{\partial \theta_j} \cdot \frac{\partial \ln p(x, \theta)}{\partial \theta_k} \right].$$

Also, however, the metric could be given by the observed information and the skewness by the observed skewness to include in this structure the work of Barndorff-Nielsen.

All the proofs of the following results will be found in [Lauritzen] unless otherwise stated.

2.10.2 Definition: Given a Statistical manifold (M, g, D) we shall define the tensor field given by $\tilde{D}(X, Y)$ implicitly by;

$$g(\tilde{D}(X, Y), Z) = D(X, Y, Z).$$

2.10.3. Definition: We define the α -connection of the Statistical manifold by

$$\nabla_X^{\alpha} Y = \bar{\nabla}_X Y - \frac{\alpha}{2} \tilde{D}(X, Y)$$

where $\bar{\nabla}$ is the Levi-Civita connection of the metric g and X, Y and Z are vector fields.

2.10.4. Theorem: If we define the α -connection as above then it is the unique torsion free connection which satisfies

$$(\nabla_X^{\alpha})g(Y, Z) = \alpha D(X, Y, Z)$$

We shall now define the *conjugate connection* of any affine connection ∇ .

2.10.5. Definition: We define the conjugate connection ∇^* to the connection ∇ by the formula

$$g(\nabla_X^* Y, Z) = Xg(Y, Z) - g(Y, \nabla_X Z)$$

where X, Y and Z are vector fields.

We get the following result about conjugacy;

2.10.6. Theorem: $(\nabla^*)^* = \nabla$.

If we let the operator Π_γ denote the parallel transport of a vector along the path γ with respect to the connection ∇ . Then Π_γ^* denotes the parallel transport with respect to ∇^* . The conjugate connections are dual in the sense that

$$g(\Pi_\gamma X, \Pi_\gamma^* Y) = g(X, Y)$$

If we now consider the α -connections we see that;

2.10.7. Theorem: $(\nabla^\alpha)^* = \nabla^{-\alpha}$.

Since any connection defines a curvature tensor we shall use the notation that ∇ induces the tensor R and ∇^* induces R^* . Then we have the following result.

2.10.8. Theorem: $R(X, Y, Z, W) = -R^*(X, Y, W, Z)$.

Also we have the following two corollaries

2.10.9. Corollary: The following conditions are equivalent

i) $R = R^*$

ii) $R(X, Y, Z, W) = -R(X, Y, W, Z)$

2.10.10. Corollary: The connection ∇ is flat if and only if ∇^* is.

We can now define the following two tensors;

$$\tilde{D}_1(X, Y) = \nabla_X^* Y - \nabla_X Y$$

and

$$D_1(X, Y, Z) = g(\tilde{D}_1(X, Y), Z)$$

So we also have the following equivalences.

2.10.11. Theorem: The following are equivalent;

- i) ∇^* is torsion free
- ii) D_1 is symmetric
- iii) $\bar{\nabla} = \frac{1}{2}(\nabla + \nabla^*)$

We can now view the Statistical manifold structure in a different way. Suppose that ∇ is given and ∇^* is torsion free. We can then define a family of connections as

$$\tilde{\nabla}_X^\alpha Y = \bar{\nabla}_X Y - \frac{\alpha}{2} \tilde{D}_1(X, Y).$$

2.10.12. Theorem: We have the following identities

$$(\tilde{\nabla}^\alpha)^* = \tilde{\nabla}^{-\alpha}, \tilde{\nabla}^1 = \nabla \text{ and } \tilde{\nabla}^{-1} = \nabla^*.$$

Also,

$$g(\nabla_X^\alpha Y, Z) - g(\nabla_Y^\alpha X, Z) = g(\bar{\nabla}_X Y, Z) - g(\bar{\nabla}_Y X, Z).$$

We define the following tensor F by

$$F(X, Y, Z, W) = (\bar{\nabla}_X D)(Y, Z, W).$$

and then we get,

2.10.13. Theorem: The following are equivalent

i) $R^\alpha = R^{-\alpha}$ for all $\alpha \in \mathbb{R}$.

ii) F is symmetric.

2.10.14. Definition: We shall define a Statistical manifold which has the above property to be a *conjugate symmetric* manifold.

These manifolds form an important subclass of the set of Statistical manifolds and the following theorem gives us an easy way to classify them.

2.10.15 Theorem: The following are sufficient for a statistical manifold to be conjugate symmetric

(A) there exists $\alpha \neq 0$ such that $R^\alpha = 0$

i.e. the manifold is α -flat for some α .

(B) there exists $\alpha \neq 0$ such that $R^\alpha = R^{-\alpha}$

As we have seen the exponential families are 1-flat so this important class is in the set of conjugate symmetric manifolds.

There is another example of where knowledge of the behaviour of the geometry for one or two values of α is sufficient to understand behaviour for all α .

2.10.16. Theorem: A regular submanifold N is totally geodesic with respect to the α -connection for all α if and only if there exist two distinct α_1 and α_2 such that N is totally geodesic with respect to the α_1 and the α_2 connections.

This property that the structure is determined by two values of α is not a general one however as the work in Chapter 4 shows.

Lauritzen also looks at the geometry of particular models which we see in Chapter 4 as well.

2.11 Conclusion.

In this chapter we have considered expected geometry from two points of view. The theory of Lauritzen and the applications of Amari. The applications we have seen fall into three types. Firstly the analysis of asymptotic expansions, secondly the work on divergence functions and thirdly, applications to reparametrisation theory. We shall see these areas in a geometric framework again in Chapter 5. For the next two chapters however we shall follow the theoretical path of Lauritzen.

Chapter Three.

Characterisation of Statistical Manifolds

3.1 Introduction.

As we have seen in Chapter 2 Lauritzen has introduced the concept of a *Statistical Manifold*. Towards the end of his paper he suggested that there were a number of open theoretical questions concerning the geometry of these manifolds which were of interest. This and the next chapter discuss a couple of these questions.

Lauritzen asks the question of how to extend to statistical manifolds the classical result from Riemannian Geometry that states that, locally, the curvature determines the metric. He points out the example that although the Gaussian and Inverse Gaussian distributions when looked at as Statistical manifolds, have the same α -curvature for each α , they are not, in fact, isomorphic even locally. From this example it is clear that the α -curvature is not a sufficiently strong indicator to characterise Statistical manifolds locally. To extend the classical result to our case we will have to look for a stronger invariant.

3.2 Classical Results.

First it is helpful to look at the classical results which we wish to extend. We quote here from [Spivak]. In the introduction they are only stated informally and it should be pointed out that the standard results are a little more complicated than they first appear.

In this theorem X is a vector field on M , V is a 2-dimensional subspace of the tangent space at p , which is written as TM_p , and $L(X,V)$ is the sectional curvature of the parallel translate of V carried along the vector field X by the exponential map $t \rightarrow \exp tX$. We recall that if R is the Riemann-Christoffel curvature tensor and v_j, u_i are two tangent vectors. The sectional curvature is defined to be

$$k(u, v) = \frac{R(u, v, u, v)}{\|u, v\|}$$

where $\| \cdot \|$ denote the area of the parallelogram defined by the vectors u, v . This curvature is well known to depend only on the plane defined by u and v (see [Dodson & Poston]).

3.2.1. Theorem Let (M, g) and (M', g') be two Riemannian manifolds, and let

$$T: TM_p \rightarrow TM'_{p'},$$

be an isometry for some $p \in M$ and some $p' \in M'$.

Suppose that

$$L(X, V) = L(T(X), T(V))$$

for all 2-dimensional subspaces $V \subset TM_p$ and all sufficiently small X . Then there is an isometry from one neighbourhood of $p \in M$ to a neighbourhood of $p' \in M'$.

We can see that this result tells us, under the above circumstances, that the two manifolds are isometric. However the proof also tells us exactly what this isometry is. If the isometry between the tangent spaces is given by $\phi: TM \rightarrow TM'$ then this extends to a local isometry between the manifolds given by

$$\Phi = (\exp_{p'}) \circ (\phi) \circ (\exp_p)^{-1}: M \rightarrow M'.$$

It is important to understand that the theorem does not say that any map between manifolds which preserves sectional curvature is an isometry, and indeed Spivak points out that this is not true in general.

This is the result for Riemannian manifolds, i.e. manifolds with a metric, we also will want to look at a more general result for a manifold with any connection. We quote here the local version of the Cartan-Ambrose-Hicks Theorem (see [Wolf])

3.2.2 Theorem: Let M and M' be manifolds having the same dimension, n . Let each have a connection on its frame bundle. Let $x \in M$ and $x' \in M'$ and choose a linear isomorphism

$$\phi: TM_x \rightarrow TM_{x'}.$$

Let U and U' be normal coordinate neighbourhoods corresponding under ϕ . We shall let

$$\Phi: U \rightarrow U'$$

be the diffeomorphism defined by $(\exp_{p'}) \circ (\phi) \circ (\exp_p)^{-1}$.

For every $z \in U$ let

$$\phi_z: TM_z \rightarrow TM'_{\Phi(z)}$$

be the map defined by $\tau' \cdot \phi \cdot \tau^{-1}$, where τ and τ' are the parallel displacement maps corresponding to translation along radial geodesics. Let R, T and R', T' denote the curvature and torsion tensors for M and M' .

Suppose for every $z \in U$ that ϕ_z sends R_z to $R'_{\Phi(z)}$ and T_z to $T'_{\Phi(z)}$. Then

$$\Phi: U \rightarrow U'$$

is an affine diffeomorphism.

$$\Phi*: M_z \rightarrow M'_{\Phi(z)}$$

is just ϕ_z , furthermore Φ is the only affine diffeomorphism $U \rightarrow U'$ which induces ϕ on M_x .

In other words if the map which we previously called

$$\Phi = (\exp_{p'}) \circ (\phi) \circ (\exp_p)^{-1}$$

is one which preserves the curvature and torsion tensors then it is an affine diffeomorphism, that is to say it preserves the connection. Furthermore, in some sense, it is unique in this respect.

We can now look to see what we must do to extend such results to the case of Statistical manifolds.

3.3 Extension to Statistical manifolds.

We recall one of the definitions in Chapter 2.

3.3.1. Definition: A Statistical manifold (M, g, ∇^{+1}) is a manifold with a metric g , and a torsion free connection ∇^{+1} whose conjugate connection ∇^{-1} is also torsion free.

3.3.2. Definition: We shall define an isomorphism between Statistical manifolds to be a map $\Psi: (M, g, T) \rightarrow (M', g', T')$ which preserves both the metric and the skewness tensor.

3.3.3. Note: This is clearly the same as the condition that

$$\Psi: (M, \nabla^\alpha) \rightarrow (M', \hat{\nabla}^\alpha)$$

preserves the α -connection for all α .

We cannot directly extend the previous theorems to Statistical manifolds by using the α -curvature as our invariant. Although, for each α , we can find a map Φ^α which preserves the α -connection we need one single map which preserves it for all α , instead of a family of affine diffeomorphisms.

Although the α -curvature isn't the invariant we are looking for, we can define another invariant related to it which will work. To each Statistical manifold we shall define an associated manifold with connection in which we shall combine all the information of the one parameter family of connections into a single connection on a higher dimensional manifold.

3.3.4. Definition: If (M, g, ∇^α) is a Statistical manifold, we define (M^+, ∇^+) to be the manifold $M^+ = M \times \mathbb{R}$.

3.3.5 Definition: We will define a metric on our manifold M^+ by;

$$g^+_{ij} = \left[\begin{array}{c|c} g_{ij} & 0 \\ \hline 0 & 1 \end{array} \right]$$

where g_{ij} is the metric on M with respect to a set of coordinates (e_i) ($i=1$ to n) and we are using the coordinates on M^+ ($=M \times \mathbb{R}$) given by

$$(e_1, e_2, \dots, e_n, \alpha)$$

where α any real number.

On this Riemannian manifold we put the connection defined as follows:

Pick some coordinate system for M , (e_i) ($i=1$ to n) say. Then with respect to this coordinate system we write down the Christoffel symbols for each of the α -connections. Let us denote these by $\Gamma^{\alpha k}_{ij}$.

We then define the connection ∇^+ by writing down its Christoffel symbols with respect to the coordinate system on M^+ given by

$$(e_1, e_2, \dots, e_n, \alpha).$$

These symbols are;

$$\Gamma^{+k}_{ij}(e_1, e_2, \dots, e_n, \alpha) = \Gamma^{\alpha k}_{ij}(e_1, e_2, \dots, e_n) \quad \text{for } 1 \leq i, j, k \leq n$$

$$\Gamma^{+k}_{ij}(e_1, e_2, \dots, e_n, \alpha) = 0 \quad \text{if any of } i, j, k = n+1.$$

Thus the connection has been defined so that its restriction on the submanifolds given by $M \times \{\alpha\}$ is the precisely the α -connection. We shall therefore try to study the Statistical manifold structure by studying the single connection ∇^+ .

3.3.6. Lemma: (i) ∇^+ is a natural connection in the sense that it does not depend on the choice of basis for M used in the definition.

(ii) ∇^+ is a torsion free connection.

Proof: (i) To show that ∇^+ is a natural connection we will change the basis on M and show that the Christoffel symbols for ∇^+ transform in the correct way.

If our new basis for M is given by (\hat{e}_i) then

$$\hat{\Gamma}^{\alpha r}_{pq} = \sum_{ijk=1}^n \Gamma^{\alpha k}_{ij} \cdot \frac{\partial e_i}{\partial \hat{e}_p} \cdot \frac{\partial e_j}{\partial \hat{e}_q} \cdot \frac{\partial \hat{e}_r}{\partial e_k} + \sum_{\mu=1}^n \frac{\partial^2 e_\mu}{\partial \hat{e}_p \partial \hat{e}_q} \cdot \frac{\partial \hat{e}_r}{\partial e_\mu}$$

So as the old coordinate system for M^+ is $(e_1, e_2, \dots, e_n, \alpha)$, the new coordinate system is given by $(\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n, \alpha)$. Therefore we need to check that

$$\hat{\Gamma}^{+r}_{pq} = \sum_{ijk=1}^{n+1} \Gamma^{+k}_{ij} \cdot \frac{\partial e_i}{\partial \hat{e}_p} \cdot \frac{\partial e_j}{\partial \hat{e}_q} \cdot \frac{\partial \hat{e}_r}{\partial e_k} + \sum_{\mu=1}^{n+1} \frac{\partial^2 e_\mu}{\partial \hat{e}_p \partial \hat{e}_q} \cdot \frac{\partial \hat{e}_r}{\partial e_\mu}$$

We shall go through the various possibilities for the values of p, q , and r to check that the required formula holds. We shall take two cases.

(a) For $1 \leq p, q, r \leq n$ the result follows because

$$\Gamma^{+n+1}_{ij} = \Gamma^{+k}_{n+1j} = \Gamma^{+k}_{in+1} = 0$$

$$\text{and } \frac{\partial e_i}{\partial \alpha} = 0 \quad (1 \leq i \leq n).$$

For the case when $r=n+1$ and any value for p and q we have

$$\frac{\partial \alpha}{\partial e_i} = 0 \quad \text{for } 1 \leq i \leq n, \text{ and } \Gamma^{+n+1}_{ij} = 0$$

therefore the first summand is zero.

We see that the second summand will equal $\frac{\partial^2 e_\mu}{\partial \hat{e}_p \partial \hat{e}_q}$ where

$\mu=n+1$ because $\frac{\partial \alpha}{\partial e_i} = 0$ for $1 \leq i \leq n$. However $\frac{\partial^2 e_\mu}{\partial \hat{e}_p \partial \hat{e}_q}$ will be

zero for p or q not equal to $n+1$ since α is independent of the other e_p 's and for $p=q=\mu=n+1$ it is also clearly zero.

(b) For the case where p (or q) is equal to $n+1$ we have that

$$\begin{aligned} \hat{\Gamma}^{+r}_{pq} &= \sum_{ijk=1}^{n+1} \Gamma^{+k}_{ij} \cdot \frac{\partial e_i}{\partial \hat{e}_p} \cdot \frac{\partial e_j}{\partial \hat{e}_q} \cdot \frac{\partial \hat{e}_r}{\partial e_k} + \sum_{\mu=1}^{n+1} \frac{\partial^2 e_\mu}{\partial \hat{e}_p \partial \hat{e}_q} \cdot \frac{\partial \hat{e}_r}{\partial e_\mu} \\ &= \sum_{jk=1}^{n+1} \Gamma^{+k}_{n+1j} \cdot 1 \cdot \frac{\partial e_j}{\partial \hat{e}_q} \cdot \frac{\partial \hat{e}_r}{\partial e_k} \\ &= 0. \end{aligned}$$

(ii) The lack of torsion on ∇^+ comes from the symmetry of the lower indices in its Christoffel symbols. This in turn is because all the α -connections are themselves torsion-free.

Q.E.D.

We shall now look at the relationship between the geodesics of our new extended connection and the geodesics of each of the α -connections.

3.3.7 Lemma: If $\gamma(t)$ is an α -geodesic then $(\gamma(t), \alpha)$ is a ∇^+ geodesic. Also all curves of the form (m_i, t) , where $(m_i) \in M$ is a constant, are geodesics.

Proof: Since $\gamma(t)$ is an α -geodesic we know for $1 \leq k \leq n$ and if t is arc length, then

$$\frac{d^2 \gamma^k}{dt^2} + \sum_{ij=1}^n \Gamma^{\alpha k}_{ij} \cdot \frac{d\gamma^i}{dt} \cdot \frac{d\gamma^j}{dt} = 0$$

So since

$$\Gamma^{+k}_{n+1j} = \Gamma^{+k}_{in+1} = 0$$

we have for $1 \leq k \leq n$

$$\frac{d^2 \gamma^k}{dt^2} + \sum_{ij=1}^{n+1} \Gamma^{+k}_{ij} \cdot \frac{d\gamma^i}{dt} \cdot \frac{d\gamma^j}{dt} = 0$$

and for $k=n+1$ since $\frac{d^2 \alpha}{dt^2} = 0$ and $\Gamma^{+n+1}_{ij} = 0$ for all i and j , we have that the curve $(\gamma(t), \alpha)$ is a ∇^+ -geodesic.

Similarly (m_j, t) is a geodesic since the geodesic equations hold; for $1 \leq k \leq n$,

$$0 + \sum_{ij=1}^{n+1} \Gamma^{+k}_{ij} \cdot \frac{d\gamma^i}{dt} \cdot \frac{d\gamma^j}{dt} = \sum_{ij=1}^n \Gamma^{+k}_{ij} \cdot 0 \cdot 0 + \Gamma^{+k}_{n+1n+1} \cdot 1 \cdot 1 = 0$$

for $k=n+1$, since $\frac{d^2 t}{dt^2} = 0$ and $\Gamma^{+n+1}_{ij} = 0$.

Q.E.D.

3.3.8. Definition: We shall call geodesics of the form $(\gamma(t), \alpha)$ *horizontal geodesics* and those of the form (m_j, t) , *vertical geodesics*.

We are now in a position to extend to Statistical manifolds the earlier classical results.

3.3.9. Theorem: Let (M_1, g_1, T_1) and (M_2, g_2, T_2) be two Statistical manifolds, and (M_1^+, ∇_1^+) , (M_2^+, ∇_2^+) the corresponding manifolds with connections. If

$$\phi: TM_1^+(m_1, 0) \rightarrow TM_2^+(m_2, 0)$$

is an isometry which respects the vertical and horizontal components of M_i^+ , and if parallel transport preserves the curvature tensor of M^+ , then ϕ can be extended to a local isomorphism of Statistical manifolds.

Proof. First of all we note that since ϕ respects the horizontal components of M_i^+ at m_i we can look at it as a map from TM_1 to TM_2 . Hence (M_1, g_1) and (M_2, g_2) are locally isometric by the map

$$(\exp_p')(\phi)(\exp_p)^{-1} = \Phi_0$$

i.e. where Φ_0 is defined through the diagram

$$\begin{array}{ccc} & \phi & \\ TM_1 & \longrightarrow & TM_2 \\ \exp \downarrow & & \exp \downarrow \\ M_1 & \longrightarrow & M_2 \end{array}$$

The exp maps are on the Riemannian manifolds (M_1, g_1) and (M_2, g_2) . This result comes from Theorem 3.2.1

Now Φ_0 sends 0-geodesics to 0-geodesics and so by Lemma 3.3.7, looked at as a map between M_1^+ to M_2^+ , it sends the ∇_1^+ -geodesics in $M_1^+|_{\alpha=0}$ to the ∇_2^+ -geodesics in $M_2^+|_{\alpha=0}$.

Further by Theorem 3.2.2, since torsion is always zero, ϕ also extends to a map from M_1^+ to M_2^+ , which we shall call Φ . This is an affine diffeomorphism. So that the geodesic of the form (m_1, t) will go to the geodesic in M_2^+ of the form (m_2, t) ,

It is clear that because Φ_0 and Φ are of the same form i.e.

$$(\exp_p')(\phi)(\exp_p)^{-1}$$

then the relationship between them is given by

$$\Phi_0 = \Phi|_{M_1^+|_{(\alpha=0)}}$$

i.e. Φ_0 is just Φ restricted to $M^+_1|_{\alpha=0} (\equiv (M_1, g_1))$.

We shall show that all vertical geodesics in M^+_1 are sent by Φ to vertical geodesics in M^+_2 .

By the uniqueness of a geodesic with a particular initial tangent direction (see [Spivak]) it is enough to show that the tangent vector in the vertical direction at $(m, 0)$, which is

$$\frac{\partial}{\partial \alpha}(m, 0)$$

is sent to the tangent vector in the vertical direction in M^+_2 which is based at $(\Phi_0(m), 0)$.

We compare the tangent space at $(m, 0)$ with that at $(m_1, 0)$. To do this we pick a basis at $(m_1, 0)$ which contains

$$\frac{\partial}{\partial \alpha}(m_1, 0)$$

$\{B\}$ say, and parallel-transport this basis to $(m, 0)$ along the geodesic, $\alpha(t)$, connecting them. (We are of course only working in a local neighbourhood of $(m_1, 0)$ so this is always possible.)

Now since Φ is defined to be $(\exp_p')(\phi)(\exp_p)^{-1}$ we see that the corresponding basis in $M^+_2|_{\alpha=0}$ i.e. the parallel transport of $\phi(\{B\})$ along the geodesic $\Phi(\alpha(t))$ will be preserved under the map Φ .

So at $(m_1, 0)$ and $(\Phi_0(m_1), 0)$ we have corresponding bases under Φ and we need to know the coordinates of the vertical tangent vectors,

$$\frac{\partial}{\partial \alpha}(m_1, 0), \quad \frac{\partial}{\partial \alpha}(\Phi_0(m_1), 0)$$

in both cases. We get these coordinates by calculating the covariant derivative of the vector field $\frac{\partial}{\partial \alpha}$ along the geodesic $\alpha(t)$ and $\Phi(\alpha(t))$.

If the coordinates are given by $(\gamma^1, \gamma^2, \dots, \gamma^{n+1})$, then the covariant derivative is given by

$$\frac{d^2\gamma^k}{dt^2} + \sum_{ij=1}^{n+1} \Gamma_{ij}^{+k} \cdot \frac{d\gamma^i}{dt} \cdot \frac{d\gamma^j}{dt} \dots (1)$$

We claim that everything in this expression is preserved by Φ .

For $1 \leq k \leq n$, Γ_{ij}^{+k} is preserved because Φ is an isometry. For any other i,j,k $\Gamma_{ij}^{+k} = 0$ by construction thus is also preserved.

So far we have shown that Φ sends $M^+_1|_{\alpha=0}$ to $M^+_2|_{\alpha=0}$ and vertical geodesics to vertical geodesics.

In the same way we show that Φ sends the horizontal subspace $M^+_1|_{\alpha=\lambda}$ to horizontal subspace $M^+_2|_{\alpha=\lambda}$ for all λ .

We shall use the fact that by lemma 3.3.8 both $M^+_1|_{\alpha=1}$ and $M^+_2|_{\alpha=1}$ are totally geodesic and generated locally by the geodesics starting at a single point. Again by the local uniqueness of geodesics it is enough to show that the map $T\Phi$ respects the horizontal decomposition of M^+_1 and M^+_2 . That is, it sends $TM^+_1|_{\alpha=\mu}$ to $TM^+_2|_{\alpha=\mu}$.

We use the same method as before by constructing two corresponding bases for $TM^+_1|_{\alpha=\mu}$ and $TM^+_2|_{\alpha=\mu}$ and parallel transport the tangent spaces at $\alpha=0$, where we have the correct decomposition. Again since the coordinates of any vector can be calculated with respect to these bases by equation (1), and because of Theorem 3.2.2, Φ is an affine diffeomorphism. As everything in equation (1) is preserved we get the result.

Define Φ_v to be Φ restricted to $M^+_1|_{\alpha=v}$, then by above

$$\Phi_v : M^+_1|_{\alpha=v} \rightarrow TM^+_2|_{\alpha=v}$$

and is of the form $(\exp_p)(\phi)(\exp_p)^{-1}$ which is for each v is an affine diffeomorphism.

Thus,

$$\Phi_{\alpha} : (M_1, \nabla_1^{\alpha}) \rightarrow (M_2, \nabla_2^{\alpha})$$

is an affine diffeomorphism for each α .

Let $\Phi_{\alpha}(x,\alpha)=(a,\alpha)$ and $\Phi_{\beta}(x,\beta)=(b,\beta)$. Since vertical geodesics are preserved the line (a,t) is sent to (a',t) . It follows that $a=b$. In other words $\Phi_{\alpha}(a,\alpha)=(\Psi(a),\alpha)$ where Ψ is independent of α .

Hence we have a single affine diffeomorphism

$$\Psi : (M_1, \nabla_1^{\alpha}) \rightarrow (M_2, \nabla_2^{\alpha})$$

which is independent of α , i.e. Ψ is locally an isomorphism of statistical manifolds. Q.E.D.

We have therefore found a solution to the problem of Lauritzen. We can see that it is not the α -curvature which characterises Statistical manifolds rather it is the curvature of the associated manifold M^+ .

3.4 Conjugate Symmetric Spaces.

We shall now consider the special case of what Lauritzen called conjugate symmetric spaces (see Chapter 2). In these spaces he shows that, as in the case of Riemannian connections, the total curvature is determined by the sectional curvature. Examples of conjugate symmetric spaces are very common and include the exponential family of distributions. First we must set up a little structure.

3.4.1 Definition: We define a metric on our manifold M^+ . With respect to our usual coordinate system let the metric g^+ be defined by;

$$g^+_{ij} = \begin{bmatrix} g_{ij} & 0 \\ 0 & 1 \end{bmatrix}$$

where g_{ij} is the metric on M .

3.4.2. Lemma: If we have a tensor on M^+ which restricts to a tensor

on one of the submanifolds given by $M^+|_{\alpha=\mu}$ then raising (or lowering) indices using g^+ and then restricting to the submanifold is compatible to using g on the restricted tensor to raise the corresponding indices.

Proof: This is an easy calculation using the fact that g_{ij} and g^{ij} are zero if either i or j is equal to $n+1$

Q.E.D.

3.4.3. Theorem: If (M, g, T) is a conjugate symmetric Statistical manifold then the curvature of the associated manifold (M^+, ∇^+) is determined by its sectional curvature.

Proof: If the curvature tensor is given by

$$R_{ijkl}^+ \quad \text{for } 1 \leq i, j, k, l \leq n+1$$

with respect to the usual coordinate system, (e_i, α) , then by a well known result (see [Spivak]) all we need to prove is that

$$R_{ijkl}^+ = -R_{jikl}^+$$

Since (M, g, T) is conjugate symmetric we know, from [Lauritzen], that

$$R_{ijkl}^\alpha = -R_{jikl}^\alpha \quad 1 \leq i, j, k, l \leq n \text{ for all } \alpha.$$

To calculate R_{ijkl}^+ we use the formula;

$$R_{ijkm}^+ = g_{sm}^+ \left(\frac{\partial}{\partial e_i} (\Gamma_{jk}^{+s}) - \frac{\partial}{\partial e_j} (\Gamma_{ik}^{+s}) \right) + \left(\Gamma_{im}^+ \cdot \Gamma_{jk}^{+r} - \Gamma_{jm}^+ \cdot \Gamma_{ik}^{+r} \right) \dots (2)$$

where $\Gamma_{ijk}^+ = \Gamma_{ij}^{+s} g_{sk}^+$.

For $1 \leq i, j, k, m \leq n$ we have that,

$$R_{ijkm}^+(e_i, \alpha) = R_{ijkm}^\alpha(e_i)$$

since $g_{n+1m}^+ = 0$ and $\Gamma_{ij}^{+n+1} = 0$.

For the case where either k or $m = n+1$ we find that

$$R_{ijkl}^+(e_i, \alpha) = 0$$

This follows from equation (2) since both $\Gamma_{jn+1}^{+r} = 0$ and $\Gamma_{im+1}^+ = 0$

For the case where either i or $j = n+1$, again we use equation (2) and find that if $i=n+1$, then for $1 \leq j, k, m \leq n$.

$$\begin{aligned} R_{ijkl}^+(e_i, \alpha) &= g_{sm}^+ \left(\frac{\partial}{\partial \alpha} \left(\Gamma_{jk}^{+s} \right) \right) \\ &= \frac{\partial}{\partial \alpha} \left(g_{sm} \cdot \Gamma_{jk}^{+s} \right) \\ &= \frac{\partial}{\partial \alpha} \left(\Gamma_{jkm}^+ \right) \\ &= \frac{\partial}{\partial \alpha} \left(\Gamma_{jkm}^0 - \frac{1}{2} \cdot T_{jkm} \right) \\ &= -\frac{1}{2} \cdot T_{jkm} \end{aligned}$$

(For any of j k or m equal to $n+1$ we have $R_{ijkl}^+(e_i, \alpha) = 0$.)

If on the other hand $j=n+1$ the by the same calculation we get

$$R_{ijkl}^+(e_i, \alpha) = \begin{matrix} 1/2 \cdot T_{ikm} & 1 \leq i, k, m \leq n \\ 0 & \text{otherwise.} \end{matrix}$$

Hence we have proved the required skew-symmetry in all cases, which completes the proof.

Q.E.D.

3.4.4. Note: The proof of this result also gives us a nice geometric view of the skewness tensor. It is simply one component of the curvature tensor for our extended manifold M^+ . Notice also that this is true for any Statistical manifold, we do not need conjugate symmetry.

We can now go back to our original motivating example of the Gaussian and Inverse Gaussian manifolds. We prove that they are not isomorphic despite having the same α -curvature. Since they are both conjugate symmetric spaces we only need to check that the sectional curvature tensors of their associated manifolds are not preserved by the map Φ , (see Theorem 3.3.9 for definition of Φ).

3.4.5 Example: The Gaussian manifold is given by the family of normal distributions $N(\mu, \sigma^2)$ i.e.

$$\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \sigma > 0$$

with respect to Lebesgue measure on \mathbf{R} . Following Lauritzen, we shall

work in the (μ, σ^2) coordinate system. The metric is given by

$$g = \frac{1}{\sigma^2} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

and the α -connection is given by the Christoffel symbols;

$$\Gamma_{11}^{\alpha 1} = \Gamma_{12}^{\alpha 2} = \Gamma_{21}^{\alpha 2} = \Gamma_{22}^{\alpha 1} = 0$$

$$\Gamma_{11}^{\alpha 2} = \frac{(1-\alpha)}{2\sigma}, \quad \Gamma_{12}^{\alpha 1} = \Gamma_{21}^{\alpha 1} = -\frac{(1+\alpha)}{\sigma}$$

$$\Gamma_{22}^{\alpha 2} = -\frac{(1+2\alpha)}{\sigma}$$

The skewness tensor is

$$\begin{aligned} T_{111} &= T_{122} = T_{212} = T_{221} = 0 \\ T_{112} &= T_{121} = T_{211} = 2/\sigma^2, \quad T_{222} = 8/\sigma^3, \end{aligned}$$

and the scalar curvature is

$$K^{\alpha}(\beta_{12}) = -(1-\alpha^2)/2$$

The inverse Gaussian distribution manifold is given by the densities;

$$x^{-\frac{3}{2}} \sqrt{\frac{\chi}{2\pi}} \exp\left(\sqrt{\chi\psi} - \frac{1}{2} \cdot (\chi x^{-1} + \psi x)\right) \quad \chi, \psi > 0$$

Again following Lauritzen we use the parametrisation given by

$$\eta = \chi^{-1} \quad \text{and} \quad \theta = \sqrt{\frac{\psi}{\chi}} \quad . \quad \text{The metric is given by}$$

$$g = \begin{bmatrix} \frac{1}{2\eta^2} & 0 \\ 0 & \frac{1}{\theta\eta} \end{bmatrix}$$

The α -connection and skewness tensor are

$$\Gamma^{\alpha 1}_{11} = -\frac{(1+\alpha)}{\eta}, \Gamma^{\alpha 2}_{11} = \Gamma^{\alpha 1}_{12} = \Gamma^{\alpha 1}_{21} = 0$$

$$\Gamma^{\alpha 2}_{12} = \Gamma^{\alpha 2}_{21} = -\frac{(1+\alpha)}{2\eta}, \Gamma^{\alpha 1}_{22} = \frac{1-\alpha}{\theta}, \Gamma^{\alpha 2}_{22} = \frac{(3\alpha-1)}{2\theta},$$

$$T_{112} = 0, T_{111} = \eta - 3, T_{122} = \frac{1}{\theta\eta^2}, T_{222} = -\frac{3}{\theta^2\eta}.$$

So we get that the sectional curvature is

$$K^{\alpha}(\sigma_{12}) = -(1-\alpha^2)/2.$$

Lauritzen points out that the map $\mu = \sqrt{2\theta}, \sigma^2 = \eta/2$ is a Riemannian isometry. So, calling this map Ψ we know from Theorem 3.3.9 that if there were an affine diffeomorphism between the two associated manifolds then it would be of the form

$$((m), \alpha) \rightarrow (\Phi(m), \alpha),$$

This map would preserve the ∇^+ -curvature tensor. The vertical component of this is given by the skewness, and we see that this is not preserved by our map Φ . Thus we find we do not have an affine diffeomorphism. This means that the two Statistical manifolds are in fact not isomorphic.

3.5 The inverse problem.

The natural question which now arises is, when you have a manifold with a connection, is it the associated manifold to a statistical one? We can now answer this with the following theorem.

3.5.1. Theorem: If (M^+, ∇^+) is a trivial R -bundle over (M, g) , an n -dimensional Riemannian manifold, such that

- (i) $M^+_{|\alpha=\text{constant}}$ are totally geodesic submanifolds for all α
- (ii) With respect to any coordinate system of the form (e_i, α) , where (e_i) is a coordinate system for (M, g) , we have the identities that $\Gamma^+_{ij}{}^k = 0$ if any of i, j or $k = n+1$.
- (iii) The component of the curvature tensor R_{n+1jkm} is a symmetric three tensor independent of α .
- (iv) At $\alpha=0$, ∇^+ restricted to the $\alpha=0$ submanifold is the Levi-Civita connection for the metric g .

Then locally the manifold (M^+, ∇^+) is the associated manifold to some unique Statistical manifold.

Notes: (1) Lemma 3.3.6 tells us that (ii) is a natural geometric condition and does not depend on which coordinate system is being used on (M, g) .

(2) We shall denote the restriction of ∇^+ to the submanifold given by $\alpha=\text{constant}$ by ∇^α . Condition (i) tells us that this is well defined and the geodesics of ∇^+ and ∇^α agree.

(3) Condition (iii) is essentially setting the value of the skewness tensor in the underlying Statistical manifold. See also Note 3.4.4.

Proof: We shall show that if we define the tensor T_{jkm} to be equal to

$$-2R_{n+1 jkm}$$

in our usual coordinate system then (M, g, T) is a Statistical manifold and M^+ is the associated manifold.

By (iii) if we define T to be the tensor with coefficients T_{ijk} in our coordinate system it is a symmetric 3 covariant tensor. Hence (M, g, T) is a Statistical manifold.

We shall work out what the associated manifold with connection is for this Statistical manifold. We calculate the α -connection for (M, g, T) by using the formula

$$g(\nabla_X^\alpha Y, Z) = g(\nabla_X^0 Y, Z) - \frac{\alpha}{2} T(X, Y, Z)$$

where ∇^0 is the Levi-Civita connection for the metric g . Hence from this we can calculate the Christoffel symbols for the α -connection to be

$$\Gamma_{ijk}^\alpha = \Gamma_{ijk}^0 - \frac{\alpha}{2} T_{ijk}$$

The Christoffel symbols for the connection on the associated manifold will be

$$\Gamma_{ij}^{+k}(e_r, \alpha) = \begin{cases} \Gamma_{ij}^{\alpha k}(e_r) & 1 \leq i, j, k \leq n \\ 0 & \text{otherwise} \end{cases}$$

Note that we raise and lower indices on tensors in the usual way using the metric on the associated manifold. See Lemma 3.4.2.

Now let us see what the Christoffel symbols are for the ∇^+ connection on the manifold M^+ with respect to the natural coordinate system.

By (ii) we know they are zero if any of the indices are $n+1$, so we only need to work in the case when $1 \leq i, j, k \leq n$. Since g^+ is independent of α , and by (ii), we can use the calculation in Theorem 3.4.3 which shows

$$\begin{aligned}
R_{n+1jkm} &= g_{sm}^+ \left(\frac{\partial}{\partial \alpha} \left(\Gamma_{jk}^{+s} \right) \right) \\
&= \frac{\partial}{\partial \alpha} \left(g_{sm} \Gamma_{jk}^{+s} \right) \\
&= \frac{\partial}{\partial \alpha} \left(\Gamma_{jkm}^+ \right)
\end{aligned}$$

Hence $\frac{\partial}{\partial \alpha} \left(\Gamma_{jkm}^+ \right)$ is independent of α , by (iii), and in fact equals $1/2 T_{jkm}$. So we have a set of ordinary differential equations to determine Γ_{jkm}^+

$$\frac{\partial}{\partial \alpha} \left(\Gamma_{jkm}^+ \right) = T_{jkm}$$

with the initial conditions from (iv) that at $\alpha = 0$

$$\Gamma_{jkm}^+ = \Gamma_{jkm}^0.$$

We prove uniqueness by using Theorem 3.3.9 which states that if the associated manifolds are locally isomorphic then so are the corresponding Statistical manifolds.

Q.E.D.

3.6 Conclusion.

We have found the geometric quantity which characterises a Statistical manifold. It is not surprising that it is related to the curvature tensor, and it is an interesting result that the essentially new component to our invariant is the skewness of the Statistical manifold (see 3.4.4). Our method of proof was to describe a construction which enabled us to embed the Statistical manifold structure into a more classical differential geometry structure. We do this by turning the Statistical manifold into a Riemannian manifold with a single connection. This procedure would hopefully work for other related geometric questions about Statistical manifolds.

We have also solved, for completeness, the inverse problem, where we classify the conditions when we can find a Riemannian manifold which corresponds to a Statistical one.

Chapter Four

Equivalences of Statistical Manifolds

4.1 Introduction.

Lauritzen (in [Lauritzen]) introduces the concept of a *Statistical Manifold* as we have seen in Chapter 2. We continue to look at the series of questions which he posed at the end of that paper. We quote:

'Some Statistical manifolds are alike locally as well as globally. Various types of likeness seem to be of interest. Of course the full isomorphism, i.e. maps from M_1 to M_2 that preserve both the Riemannian metric and the skewness tensor. But also maps that preserve some structure, but not all of it could be of interest, in analogy with the notion of a conformal map in Riemannian geometry. There are several possibilities here. Isometries that preserve the skewness tensor up to a scalar or up to a function. Maps that preserve the metric up to scalars and do and do not preserve skewness etc.'

This chapter starts to explore the area of maps between Statistical manifolds in the context of Amari's expected geometry on exponential and curved exponential families. We look at the relevance of conformal mappings to Statistical manifolds. We see also how statistical properties of these manifolds fix their geometric structure.

4.2 Statistical Framework.

We shall in this section set up the statistical framework in which we work.

It is not sensible to study the complete set of maps between Statistical manifolds since most of these will have no statistical significance. We have restricted our attention to the most natural map

from a statistical point of view.

We look at the maximum likelihood estimator as a natural map from the sample space X to the manifold, S , of exponential distributions. We know that, with repeated sampling, the mean of the sample is a sufficient statistic. We shall use this and identify the space of repeated samples X^n with X using the mean. Further let us assume that the Fisher information matrix is everywhere nonsingular. Thus we have $\hat{\theta} : X \rightarrow S$ is locally a bijection (see below). So we have a natural correspondence between X and S . Let us suppose we have two equidimensional exponential families S_1 and S_2 with the same sample space X and their denote their m.l.e.s by

$$\hat{\theta}_i : X \rightarrow S_i \quad (i=1,2),$$

Using the natural correspondences between X and S_1 and X and S_2 , we can produce a natural correspondence between S_1 and S_2 simply by using the map

$$(\hat{\theta}_1)^{-1}(\hat{\theta}_2) = \Psi.$$

We shall call Ψ the natural map between S_1 and S_2 . For the rest of this chapter we shall assume we are working in the framework above.

We are using the maximum likelihood estimator in both cases. All the map, Ψ , does, of course, is to compare the different estimates from the same sample in two different exponential models. There are alternatives, for example we could use a different estimation procedure in each family.

The general case is going to be complicated by the fact that different models will give different sets of ancillary submanifolds in the sample space, X^n . We here are dealing with the very special case where the mean is a sufficient statistic in each model thus the ancillaries are the same.

4.3 First Order Equivalence.

4.3.1 Lemma: Ψ is a diffeomorphism.

Proof: This is an application of the Inverse Function Theorem. To prove that a smooth map between equidimensional manifolds is a

diffeomorphism it is sufficient to show that its derivative is non-singular everywhere.

We can study the derivative of Ψ by seeing it as the composite of the derivatives of the maps $(\hat{\theta}_1)^{-1}$ and $(\hat{\theta}_2)$. Thus it is sufficient to show that the derivatives of the maps $(\hat{\theta}_1)$ and $(\hat{\theta}_2)$ have full rank.

Since we are in an exponential family we know by a simple calculation that the Fisher information is equal to minus the hessian of the log-likelihood function. But this is also equal to the second derivatives of the maximum likelihood estimator. Therefore, since by assumption the Fisher information is non-singular we have the result.

Q.E.D.

A conformal map, f , between two manifolds with metrics, (M_1, g_1) and (M_2, g_2) , is one which preserves angles measured in the corresponding metrics.

From this definition we get the idea of two metrics being conformally equivalent.

4.3.2. Definition: Let v and w be any vectors in the vector space $TM_1(x)$ then g_1 and g_2 are conformally equivalent if there exists a diffeomorphism $f(x)$ from M_1 to M_2 such that for all v, w and x we have;

$$g_1(v, w) = f(x) \cdot g_2(f^*(v), f^*(w))$$

where f^* is the induced map between tangent spaces, (see [Spivak]).

This definition means that g_1 and g_2 differ by a function from M_1 to \mathbb{R} .

We can now apply this idea to Statistical manifolds. We shall quote here a result from Amari about the first order efficiency of an estimator in an exponential family. We defined this in Chapter 2. The framework for this theorem is that we have a full exponential family M with a curved exponential family N as a submanifold. Amari looks at the class of estimators on the submanifolds N as being defined by the set of ancillary manifolds which cut N transversely. He then studies the estimators by looking at the geometric properties of these ancillary

submanifolds. For more details about this see Chapter 2. We shall just quote the results we need.

4.3.3 Theorem: A consistent estimator is first order efficient when and only when the associated ancillary family is orthogonal.

4.3.4 Definition: We shall define two Statistical manifolds (M_1, g_1, T_1) and (M_2, g_2, T_2) to be first order equivalent if the natural diffeomorphism between them is such that the class of first order efficient ancillary families of (M_1, g_1, T_1) are mapped onto the class of first order efficient ancillary families of (M_2, g_2, T_2) .

Then we have the result

4.3.5 Theorem: (M_1, g_1, T_1) and (M_2, g_2, T_2) are first order equivalent (by Ψ the natural map) iff g_1 and g_2 are conformally equivalent.

Proof: Since we are dealing with the natural diffeomorphism we know that consistent estimators are mapped to consistent estimators, so that the correspondence between ancillary submanifolds and estimators of curved exponential families inside (M_1, g_1, T_1) and (M_2, g_2, T_2) are preserved.

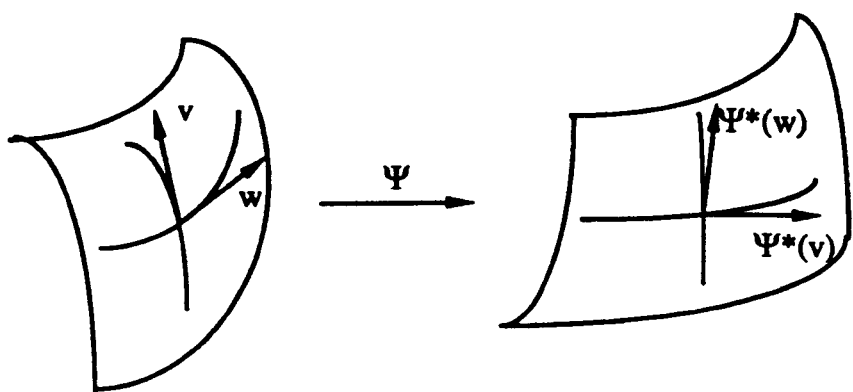
First we shall assume g_1 and g_2 are conformally equivalent by Ψ . Since angles are preserved by Ψ , orthogonal families are mapped to orthogonal families. Hence we have first order equivalence.

Conversely, if we have first order equivalence, then g_1 and g_2 agree on orthogonal pairs of tangent vectors at any point x , $\{v, w\}_x$. Thus our diffeomorphism Ψ will take an orthogonal pair and map it to an orthogonal pair. This is because given any such pair we construct the geodesics through the point x in the directions v and w . Then we can treat one of these geodesics as an hypothesis and the other as a first order efficient estimator. This will be mapped by Ψ to another first order efficient estimator whose tangent vector will be the image of w , i.e. $\Psi^*(w)$, while the image of the hypothesis will have as its tangent vector, $\Psi^*(v)$. Since the first order efficiency is preserved by Ψ we must have by Theorem 4.3.4, that $\Psi^*(w)$ is orthogonal to $\Psi^*(v)$.

4.3.6 Lemma: Up to conformal equivalence a metric is determined by its choice of orthogonal pairs.

Proof: We have shown that our two metrics g_1 and g_2 agree on orthogonal pairs, i.e.

$$g_1(v,w) = 0 \quad \text{iff} \quad g_2(\Psi^*(v), \Psi^*(w)) = 0.$$



Since the g_i are both symmetric bilinear forms it is possible to pick an orthogonal basis such that g_1 has the representation

$$\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}$$

Let this basis be $\{v_1, v_2, \dots, v_n\}$. Then, with respect to the basis $\{ \Psi^*(v_1), \Psi^*(v_2), \dots, \Psi^*(v_n) \}$ g_2 must have the representative

$$\begin{bmatrix} \mu_1 & 0 & \dots & 0 \\ 0 & \mu_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \mu_n \end{bmatrix}$$

because orthogonal pairs are preserved.
Consider the pair

$$\{(1/\sqrt{\lambda_1}).v_1 + (1/\sqrt{\lambda_2}).v_2, (1/\sqrt{\lambda_1}).v_1 - (1/\sqrt{\lambda_2}).v_2\}$$

Clearly these are g_1 orthogonal, hence g_2 orthogonal. Therefore,

$$\begin{aligned} 0 &= g_2((1/\sqrt{\lambda_1}).v_1 + (1/\sqrt{\lambda_2}).v_2, (1/\sqrt{\lambda_1}).v_1 - (1/\sqrt{\lambda_2}).v_2) \\ &= \mu_1/\lambda_1 - \mu_2/\lambda_2 \end{aligned}$$

and by the same reasoning for the other indices we find at x the matrices for g_1 and g_2 just differ by a constant multiple, $f(x)$, say.

Hence

$$g_2 = f(x).g_1,$$

or g_2 and g_1 are conformally equivalent.

Q.E.D.

This proves the lemma and hence the theorem.

Q.E.D.

4.4 Second order equivalence.

The second order equivalence problem is in fact not very interesting due to the following theorem by Amari.

4.4.1 Theorem: A bias-corrected first order efficient estimator is automatically second order asymptotic efficient.

Following Amari we shall only deal with the class of bias-corrected estimators. We define two manifolds to be second order (unbiased) efficient if the natural diffeomorphism Ψ between them is such that the class of second order efficient estimators are mapped onto the class of second order efficient estimators. Hence we have the

following easy corollary of Theorem 4.4.1.

4.4.2 Corollary: (M_1, g_1) and (M_2, g_2) are second order equivalent iff they are conformally equivalent.

4.5 Third order equivalence.

Again quoting from Amari we have the following result.

4.5.1 Theorem: An estimator is third order asymptotically efficient when and only when the associated ancillary submanifold has zero mixture curvature (see Chapter 2).

4.5.2 Definition: We shall define two manifolds to be third order equivalent if the natural diffeomorphism between them sends the class of third order efficient estimators onto the class of third order efficient estimators.

Then we have the following result connecting two Statistical manifolds. We use the notations for Statistical manifolds which we discussed in Chapter 2.

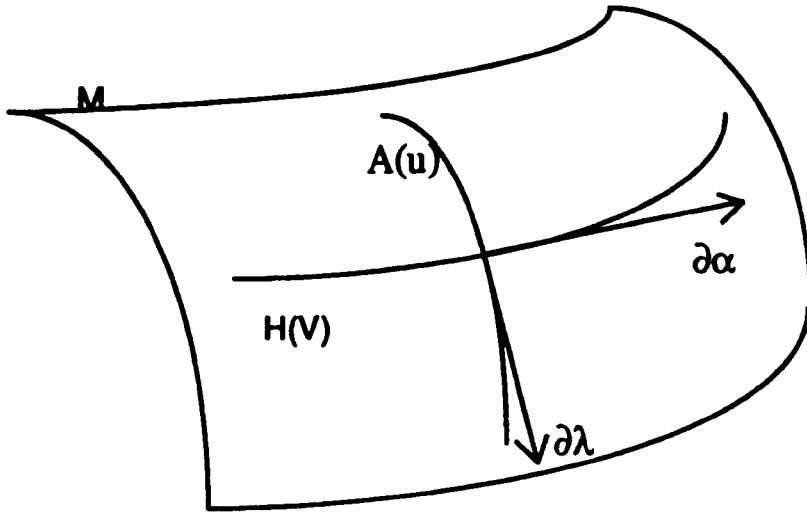
4.5.3. Theorem: Let $(M, g_1, \nabla_1^{(-1)})$ and $(M, g_2, \nabla_2^{(-1)})$ be two Statistical manifolds with conformal Fisher information metrics. Then if these two are third order equivalent then $\nabla_1^{(-1)}$ completely determines $\nabla_2^{(-1)}$.

Further if $g_1 = g_2$ then we know that $\nabla_1^{(-1)} = \nabla_2^{(-1)}$.

Note: This theorem shows that two isometric Riemannian manifolds (M, g_1) and (M, g_2) will give isomorphic Statistical manifolds $(M, g_1, \nabla_1^{(-1)})$ and $(M, g_2, \nabla_2^{(-1)})$ if and only if the Statistical manifolds are third order equivalent. However it does point out that two Statistical manifolds can be third order equivalent without being isomorphic if their metrics are only conformally equivalent rather than isometric.

Proof. By theorem 4.5.2 clearly $(M, g_1, \nabla_1^{(-1)})$ and $(M, g_2, \nabla_2^{(-1)})$ are first and second order equivalent.

Let our hypothesis in M be given by $H(v)$, and the ancillary family by $A(u)$ (see figure). Further we shall denote vector fields in the ancillary directions by $\partial\lambda_i$ and along the hypothesis submanifold by $\partial\alpha_i$.



Now the mixture curvature of $A(u)$ in M is defined to be

$$\langle \nabla_{\partial\lambda_i}^{-1} \partial\lambda_j, \partial\alpha_i \rangle$$

where \langle , \rangle denotes the metric on M . So clearly if $\nabla_1^{(-1)}$ and $\nabla_2^{(-1)}$ are the same then we have third order equivalence.

Conversely let us suppose that the two manifolds are third order equivalent then that would imply that

$$\langle (\nabla_1)^{-1}_{\partial\lambda_i} \partial\lambda_j, \partial\alpha_i \rangle = 0 \text{ iff } \langle (\nabla_2)^{-1}_{\partial\lambda_i} \partial\lambda_j, \partial\alpha_i \rangle = 0$$

for all vector fields $\partial\lambda$ and $\partial\alpha$ such that

$$\langle \partial\alpha_i, \partial\lambda_j \rangle = 0 \text{ for all } i \text{ and } j.$$

This condition implies that both $\nabla_1^{(-1)}$ and $\nabla_2^{(-1)}$ have the same autoparallel submanifolds. This is because a submanifold, N is $\nabla_1^{(-1)}$ -autoparallel if for all vector fields X and Y we have the property that $\nabla_1^{(-1)}Y$ lies in the tangent bundle of N . Now for symmetric

connections we know that autoparallel is equivalent to totally geodesic (See Spivak). Since all our connections are torsion free, and therefore symmetric we have that $\nabla_1^{(-1)}$ and $\nabla_2^{(-1)}$ have the same totally geodesic submanifolds. In particular they have the same geodesics.

To finish we note that a torsion free connection is determined by its geodesics and the parametrisation on them (see [Spivak]). We are dealing with torsion free connections thus the $\nabla_2^{(-1)}$ connection is completely determined by the $\nabla_1^{(-1)}$ connection.

Q.E.D.

4.6 Conclusion.

We have seen in this chapter some results on classifying maps between Statistical manifolds. One of the important ideas here is that we should not look at all maps between such manifolds since they will not all have statistical significance. We defined the natural map between two exponential families. This is sensible here because of the nice properties of these families. In general it would be interesting to extend this definition to a wider class of Statistical manifolds.

The results we have proved here have a useful role to play when we are considering the general theory of how to apply geometry to statistics. They show that, at least for the exponential case, a Statistical manifold is determined by third order information. Therefore to get more detail about the statistics we would have to extend the structure. In the rest of the thesis we start to study a new geometric structure which does this.

Chapter Five

Introduction to Preferred point geometry.

5.1 Introduction.

In this chapter we shall introduce the essentially new concept of this thesis. In the previous chapters we have explored the current state of statistical geometry and looked at some of the open problems in Lauritzen's Statistical manifold theory. We shall now diverge from these standard theories by introducing the idea of a *preferred point geometry*, and thereby produce several new generalisations of the current theories. We shall begin by studying some of the weaknesses of Statistical manifold theory and also some of its more unexplained and more unnatural features. By following these problems we find that we are led in a geometric way to our new structure. We should point out that while the path we follow seems to be a natural one in a geometric sense, from a statistical point of view there are questions about which is the best approach. We find that we produce several generalisations of Amari's and Barndorff-Nielsen's ideas each of which will contain certain statistical information. In later chapters we shall start to examine the statistical uses of each of our new structures.

Let us describe the work previously considered in earlier chapters as *Statistical manifold theory*. There seem to be two main questions concerning this theory as we shall now see. The first is a geometric one and the second statistical.

The first problem is the question of how natural is the Statistical manifold structure (M, g, T) (see Chapter 2) from a purely geometric viewpoint. The idea is certainly new to differential geometry. There are many examples of manifolds with metrics and their corresponding metric connections. Also there are many examples of non-metric connections being used, particularly in Theoretical Physics. What is new, however is the way in which the metric and non-metric α -connections are related. This is certainly unusual and it would be interesting to find some purely geometric examples of this structure. That is examples with no statistical information at all. This would explain whether the underlying structure is forced by statistical considerations or is more independent. The main point which needs to be explained is; what produces the one parameter family of connections? We would like to know what further geometric information we need to impose on a Riemannian manifold to produce the new structure.

Preferred Point Geometry

The second question which needs to be clarified is a purely statistical one. We have seen the uses of Statistical manifold theory in three main areas. The first of these is inference, using the geometry to measure distances on a manifold, also the related problem of finding divergence functions (see Chapter 2). The second area is that of the reparametrisation of parametric structures i.e. using particular affine coordinates to our advantage. The third is the higher order theory of asymptotic analysis, using curvature to interpret the higher order terms of asymptotic expansions. In all these areas we find the same statistical manifold structure being useful. The question is why should three essentially different types of problems give the same geometric structure? Is there some underlying principle which would unite them and which forces the geometry to Lauritzen's theory?

If we look at the geometric basis underlying these three areas we see certain problems and questions which force us towards our new geometry.

In the area of inference we have found applications to the theory of divergence functions. However there is a major difficulty here in trying to apply standard differential geometry. Many examples of divergence functions have been proposed which are fundamentally non-symmetric. That is the divergence of θ_1 to θ_2 i.e., $d(\theta_1, \theta_2)$, does not equal that of θ_2 to θ_1 . The typical example of this is the Kullback-Leibler divergence (see [Chentsov]).

$$d(\theta_1, \theta_2) = E_{\theta_1}[\ln p(x, \theta_1) - \ln p(x, \theta_2)].$$

However, almost all of classical differential geometry has a symmetric structure. The geodesic distance from θ_1 to θ_2 in any Riemannian manifold will always equal that of θ_2 to θ_1 . Indeed much of the application of classical differential geometry to theoretical physics has been based on the fact that geometric quantities are independent of the observer, i.e. all points are treated equally. Thus, maybe we should be trying to fit the inference problem into a non-symmetric differential geometry.

Another strong argument for the use of a non-symmetric geometry for the modeling of inference theory is the fundamental Neyman-Pearson Lemma which in effect recognises the fundamental asymmetry of statistics by distinguishing between *size* and *power* in inference theory. In particular we note that the power function itself is not a symmetric one.

Consider now the case of applying differential geometry to asymptotic analysis. If we study the heuristic justification for using the Fisher metric which was proposed in [Amari] we see that there is a problem. The reason for using the Fisher metric is simply that as the sample size increases in probability the

Preferred Point Geometry

maximum likelihood estimator tends to the true parameter value. Further the asymptotic distribution of the m.l.e. is normal with a variance which is the inverse of the Fisher information at the true parameter. This is a reasonable justification for using the Fisher information as a metric at the true parameter, but surely only there. There seems little justification in this argument for using the Fisher information at other points.

The implication of this remark is simply that maybe we should study a geometry which is dependent in some way on the true parameter value. In other words unlike classical differential geometry where all points are treated equally, we should perhaps be studying a geometry where one particular point is singled out from the manifold as being different. To stretch the analogy with the application of geometry into physics we should like a geometry which is dependent on the observer. This is a completely natural approach in the classical statistical context where the true parameter value has a preferred role.

5.2 Preferred point Geometry.

Following the previous rather vague discussion we shall now try to use the ideas which have come up to produce a new, purely geometric, structure. The main points we require for this structure are firstly that there must be some inherent non-symmetry reflected in the geometry, and secondly that there is some special or *preferred point*. In application to statistics this point will be maybe the true parameter or some estimate of it from the observed data. However here we shall work purely geometrically. The ideal result will be that having produced such a purely geometric system we will find that it has parallels with the Statistical manifold structure. In fact we produce a purely geometric structure which is a clear generalisation of a Statistical manifold, i.e. the Statistical manifold is a first order approximation to our new *preferred point geometry*.

Let us consider any Riemannian structure whose metric is smoothly dependent on one (preferred) point. We shall denote this by $(M, g^{\theta_0}, \theta_0)$, where M is some finite dimensional manifold, θ_0 is the preferred point in M , and g^{θ_0} is the metric whose value depends on θ_0 .

Thus we could view the whole preferred point manifold as a set of Riemannian manifolds $(M_{\theta_0}, g^{\theta_0})$ indexed by the points of the manifold itself.

Preferred Point Geometry

However in most of the following work we shall consider θ_0 fixed and work in just one point of this set.

It will be helpful to view the preferred point manifold $(M, g^{\theta_0}, \theta_0)$ in a different way. We shall take an underlying Riemannian manifold (M, g) and consider the preferred point metric as a perturbation of g , where the perturbation depends on the preferred point. Thus we shall also use the notation $(M, g, \theta_0, h^{\theta_0})$ for a preferred point manifold, where M is our manifold, g a fixed metric, θ_0 the preferred point and h^{θ_0} a symmetric 2-Tensor which depends on θ_0 and is the perturbation of g . The relationship between $(M, g^{\theta_0}, \theta_0)$ and $(M, g, \theta_0, h^{\theta_0})$ is given by the formula:

$$g^{\theta_0} = g + h^{\theta_0}$$

We shall (for the moment) assume that the perturbation is such that $g + h^{\theta_0}$ is always non-singular and positive definite.

In application we shall often have g , the underlying metric, as the Fisher information and h^{θ_0} will be a small perturbation which is zero at θ_0 . In this case we note that, since g is both non-singular and positive definite, then our assumption that this is also true for the perturbed metric will always hold in a neighbourhood of θ_0 .

We see that this structure has both the dependence on a preferred point and also the lack of symmetry which we require. The reason for this lack of symmetry is clear if we define the preferred point geodesic distance $d_{g^{\theta_0}}(\theta_0, \theta)$ to be the geodesic distance from the preferred point θ_0 to any other point θ measured in the g^{θ_0} metric. If we now reverse the role of the two points to get $d_{g^{\theta_0}}(\theta, \theta_0)$, we will get a different distance since we are now measuring in the g^{θ} -metric rather than the g^{θ_0} -metric.

We shall now do some elementary differential geometry on our preferred point manifold. First we shall calculate the Levi-Civita connection.

The Christoffel symbols Γ_{ijk} for a metric g_{ij} on a p -dimensional manifold are given by

Preferred Point Geometry

$$\Gamma_{ijk} = \frac{1}{2} \left(\frac{\partial g_{jk}}{\partial \theta_i} + \frac{\partial g_{ik}}{\partial \theta_j} - \frac{\partial g_{ij}}{\partial \theta_k} \right)$$

Thus for our preferred point metric

$$\begin{aligned} \Gamma^{\theta_0}_{ijk} &= \frac{1}{2} \left(\frac{\partial g_{jk}}{\partial \theta_i} + \frac{\partial g_{ik}}{\partial \theta_j} - \frac{\partial g_{ij}}{\partial \theta_k} \right) + \frac{1}{2} \left(\frac{\partial h^{\theta_0}_{jk}}{\partial \theta_i} + \frac{\partial h^{\theta_0}_{ik}}{\partial \theta_j} - \frac{\partial h^{\theta_0}_{ij}}{\partial \theta_k} \right) \\ &= \Gamma_{ijk} + T^{\theta_0}_{ijk} \end{aligned} \quad (1)$$

We can immediately see a parallel with Lauritzen's connection if for instance $T^{\theta_0}_{ijk}$ is the skewness tensor. However, in the above definition T^{θ_0} is not symmetric nor is it a tensor since it is the Christoffel symbol for the h^{θ_0} perturbation. To show how to recover Lauritzen's structure from equation (1) we need to consider the perturbation as being a small one and then consider the theory of asymptotic approximations on a manifold.

5.3 Power series expansions

Let us consider an expansion in a power series form, i.e.

$$g^{\theta_0}_{ij} = g_{ij} + (\theta - \theta_0)^k T_{ijk} + (\theta - \theta_0)^k (\theta - \theta_0)^l S_{ijkl} + \dots$$

where $g^{\theta_0}_{ij}$ is a tensor.

Let us see how this expression behaves under a change of coordinates

$$\theta \rightarrow \phi,$$

say. Since $g^{\theta_0}_{ij}$ is a tensor we have, in ϕ coordinates, that it is expressed as

$$\frac{\partial \theta^\alpha}{\partial \phi_i} \cdot \frac{\partial \theta^\beta}{\partial \phi_j} \cdot g^{\theta_0}_{\alpha\beta}$$

Preferred Point Geometry

Thus the expression becomes

$$\frac{\partial \theta^\alpha}{\partial \phi_i} \cdot \frac{\partial \theta^\beta}{\partial \phi_j} \cdot g_{\alpha\beta}^{\theta_0} = \frac{\partial \theta^\alpha}{\partial \phi_i} \cdot \frac{\partial \theta^\beta}{\partial \phi_j} \cdot g_{\alpha\beta} + (\theta - \theta_0)^k \frac{\partial \theta^\alpha}{\partial \phi_i} \cdot \frac{\partial \theta^\beta}{\partial \phi_j} \cdot T_{\alpha\beta k} + \dots$$

however, for consistency, we want the power series in terms of $(\phi - \phi_0)^k$. Since

$$(\theta - \theta_0) = (\phi - \phi_0)^k (\partial \theta / \partial \phi^k) + \text{higher order terms},$$

we see that the formula becomes,

$$\frac{\partial \theta^\alpha}{\partial \phi_i} \cdot \frac{\partial \theta^\beta}{\partial \phi_j} \cdot g_{\alpha\beta}^{\theta_0} = \frac{\partial \theta^\alpha}{\partial \phi_i} \cdot \frac{\partial \theta^\beta}{\partial \phi_j} \cdot g_{\alpha\beta} + (\phi - \phi_0)^k \frac{\partial \theta^\alpha}{\partial \phi_i} \cdot \frac{\partial \theta^\beta}{\partial \phi_j} \cdot \frac{\partial \theta^\lambda}{\partial \phi_k} \cdot T_{\alpha\beta\lambda} + \dots$$

Thus we see that if we look at the coefficients of the linear term before and after the coordinate transformation, i.e. T_{ijk} and $\frac{\partial \theta^\alpha}{\partial \phi_i} \cdot \frac{\partial \theta^\beta}{\partial \phi_j} \cdot \frac{\partial \theta^\lambda}{\partial \phi_k} \cdot T_{\alpha\beta\lambda}$ then this particular coefficient transforms like a tensor. Therefore the first order approximation of $g_{ij}^{\theta_0}$ is $g_{ij} + (\phi - \phi_0)^k T_{ijk}$ where T_{ijk} is a tensor. Because of this we shall consider our perturbation of the metric as a power series. Thus we set

$$h_{ij}^{\theta_0} = (\theta - \theta_0)^k T_{ijk} + \text{higher order terms}.$$

Calculating the Levi-Civita connection we see that

$$\Gamma_{ijk}^{\theta_0} = \Gamma_{ijk} + 1/2(T_{kij} + T_{kji} - T_{ijk}) + \text{first order terms}.$$

If we impose the extra condition that T_{ijk} be a symmetric 3- tensor we find that,

$$\Gamma_{ijk}^{\theta_0} = \Gamma_{ijk} + 1/2(T_{ijk}) + \text{first order terms},$$

where T is a symmetric 3-tensor, i.e., the first term of the expansion of the Christoffel symbol is precisely an α -connection in the Lauritzen sense. This means that the structure that we have developed does in fact have some of the properties of Lauritzen's. The curvature at the point is measured by a connection which is different from the Levi-Civita connection of the underlying (unperturbed) metric. Also, at least as a first approximation, the two structures agree.

We shall say that if two tensors agree up to the first term of such a power series then one is a first order approximation of the other or they are equivalent to the first order.

The above calculation raises an important point. Even though two metrics agree to the first term at a point their respective connections do not. There exists a zeroth order correction term due to the first order difference between the metrics. Thus we can see that we can view the skewness tensor of Lauritzen as this correction term.

We shall sum up the above discussion in the following theorems.

5.3.1 Theorem: Consider the perturbation of a metric g_{ij} which is a power series in some coordinate system around θ_0 .

$$g^{\theta_0}_{ij} = g_{ij} + (\theta - \theta_0)^k T_{ijk} + \text{higher order terms in } (\theta - \theta_0).$$

where T_{ijk} is a symmetric 3-tensor.

This is a geometrically well defined perturbation, i.e. the definition is independent of the coordinate system modulo the $(\theta - \theta_0)^2$ terms.

Proof. We have seen that the coefficient of the $(\theta - \theta_0)$ term transforms as a 3-tensor. Thus in any coordinate system its coefficient will be the symmetric 3-tensor T_{ijk} .

Q.E.D.

5.3.2 Theorem: If $g^{\theta_0}_{ij}$ is defined as above then the corresponding Christoffel symbols are:

$$\Gamma^{\theta_0}_{ijk} = \Gamma_{ijk} + 1/2 (T_{ijk}) + \text{first order terms}.$$

Where Γ_{ijk} are the Christoffel symbols of the unperturbed metric g .

Preferred Point Geometry

We have seen that a first order perturbation of the metric produces a zeroth order perturbation (that is a perturbation of the constant term) of the Christoffel symbols. We call this the first order correction term.

5.3.3 Corollary: The first order correction term of the Christoffel symbols is geometrically well defined and is the tensor $1/2T$.

Thus we have produced a purely geometric example of a structure which is a generalisation of Lauritzen's structure. We note that the relationship between the two structures is much stronger than just an approximation. In most of the uses of Amari's geometry we only use the geometric structures evaluated at θ_0 . If we are only interested in the geometry at this point then the two structures are the same. The reason for this is that when θ equals θ_0 then the first order approximations are exact. Thus at θ_0 the exact metric is the unperturbed g_{ij} but the exact curvature is given by $\Gamma_{ijk} + 1/2T_{ijk}$. This is not the connection which you would have expected if you only had knowledge of the unperturbed metric. Thus we have produced a purely geometric structure which as far as the theorems of Amari are concerned are actually identical. Also since we have also said that the justification for the Fisher metric is strongest at the true parameter it is not surprising that away from this point a perturbation of the Fisher metric is possible.

The following lemma shows that the first order correction behaves nicely as a tensor with respect to the operation of raising and lowering indices.

5.3.4. Lemma: Let Γ_{ijk} and $\Gamma^{\theta_0}_{ijk}$ be the Christoffel symbols for the metrics g and g^{θ_0} , where to first order

$$g^{\theta_0}_{ij} = g_{ij} + (\theta - \theta_0)^k T_{ijk}$$

Let T_{ijk} be a symmetric 3-tensor. Then the first order difference between Γ_{ijk} and $\Gamma^{\theta_0}_{ijk}$ is T_{ijk} , and the first order difference between Γ^k_{ij} and $\Gamma^{\theta_0 k}_{ij}$ is T^k_{ij} .

Preferred Point Geometry

Proof. By definition

$$\Gamma_{ij}^k = g^{kl} \cdot \Gamma_{ijl}$$

where g^{ij} is the inverse of g_{ij} .

Thus,

$$\begin{aligned} (\Gamma^{\theta_0})_{ij}^k &= (g^{\theta_0})^{kl} \cdot (\Gamma^{\theta_0})_{ijl} \\ &= (g^{kl} + (\theta - \theta_0)^s S_{kls} + \text{higher terms}) \cdot (\Gamma_{ijl} + 1/2 (T_{ijl}) + O(\theta - \theta_0)) \\ &= g^{kl} \Gamma_{ijl} + 1/2 g^{kl} T_{ijl} + O(\theta - \theta_0) \\ &= \Gamma_{ij}^k + T_{ij}^k + O(\theta - \theta_0). \end{aligned}$$

Q.E.D.

5.4 Examples.

We shall now give two examples, from statistics, of the previous ideas. We shall look at their statistical motivation and importance in later chapters. Here we shall use them as an illustration of the theory of preferred point geometries.

Consider the expression

$$g^{\theta_0}_{ij} = E_{p(x, \theta_0)} [\partial_i \ln p(x, \theta) \cdot \partial_j \ln p(x, \theta)]$$

where $\partial_i = \partial / \partial \theta_i$, and where we are working in some finite dimensional parametric system of distributions $\{ p(x, \theta) \mid \theta \in \mathbb{R}^p \}$.

We point out that the important difference between the Fisher information and g^{θ_0} . With g^{θ_0} the expectation is always taken over our fixed preferred point θ_0 whereas when using the Fisher information as a metric we use

$$g_{ij} = E_{p(x, \theta)} [\partial_i \ln p(x, \theta) \cdot \partial_j \ln p(x, \theta)]$$

From a classical statistical point of view our expression g^{θ_0} is a very natural object to take since all expectations are taken over the true parameter which will of course be fixed in any particular instance.

5.4.1 Lemma: In a neighbourhood around θ_0 , g^{θ_0} is a metric.

Proof. We see that g^{θ_0} is symmetric and transforms like a tensor. Since at θ_0 g^{θ_0} is equal to the non-singular Fisher information, then locally g^{θ_0} must also be non-singular.

Q.E.D.

We shall now see, by expanding $p(x, \theta_0)$ around θ_0 , that g^{θ_0} is a power series perturbation of the Fisher information. This is the sort of construction which was considered in the previous section.

$$g_{ij}^{\theta_0} = \int_X \frac{\partial}{\partial \theta_i} \cdot \ln p(x, \theta) \cdot \frac{\partial}{\partial \theta_j} \cdot \ln p(x, \theta) \cdot p(x, \theta_0) \cdot d\mu$$

where $\mu = \mu(x)$ is the measure on the sample space X .

Writing $\partial/\partial\theta_i$ as ∂_i we have

$$\begin{aligned} g_{ij}^{\theta_0} &= \int_X \frac{\partial}{\partial \theta_i} \cdot \ln p(x, \theta) \cdot \frac{\partial}{\partial \theta_j} \cdot \ln p(x, \theta) \cdot p(x, \theta_0) \cdot d\mu \\ &= \int_X \partial_i \ln p(x, \theta) \partial_j \ln p(x, \theta) \cdot \left(p(x, \theta) - (\theta - \theta_0)^k \frac{\partial}{\partial \theta_k} p(x, \theta) + O((\theta - \theta_0)^2) \right) \cdot d\mu \\ &= \int_X \frac{\partial}{\partial \theta_i} \cdot \ln p(x, \theta) \cdot \frac{\partial}{\partial \theta_j} \cdot \ln p(x, \theta) \cdot p(x, \theta) \cdot d\mu - \\ &\quad (\theta - \theta_0)^k \int_X \partial_i \ln p(x, \theta) \partial_j \ln p(x, \theta) \frac{\partial}{\partial \theta_k} p(x, \theta) \cdot p(x, \theta) \cdot d\mu + O((\theta - \theta_0)^2) \\ &= g_{ij} - (\theta - \theta_0)^k T_{ijk} + O((\theta - \theta_0)^2) \end{aligned}$$

Preferred Point Geometry

where $T_{ijk} = E_{\theta}[(\partial_i \ln p)(\partial_j \ln p)(\partial_k \ln p)]$ which is precisely Lauritzen's skewness tensor, (see Chapter 2).

Thus we have the conditions of theorem 5.3.2. Hence the Levi-Civita connection of g^{θ_0} is

$$\Gamma_{ijk} = 1/2 T_{ijk} + O(\theta - \theta_0).$$

Which to first order is the -1-connection of Lauritzen/Amari.

It is important to note again that while in general the Levi-Civita connection of g^{θ_0} is only first order equivalent to the -1-connection, at the preferred point θ_0 the two connections agree exactly. Thus if we are just using the connection to measure the curvature at θ_0 then the two measures agree.

There are many examples of preferred point metrics. Let us consider the following set of metrics.

$$g_{ij}^{\theta_0, \alpha} = \int_X \left(\frac{p(x, \theta)}{p(x, \theta_0)} \right)^\alpha \frac{\partial}{\partial \theta_i} \ln p(x, \theta) \cdot \frac{\partial}{\partial \theta_j} \ln p(x, \theta) \cdot p(x, \theta_0) \cdot d\mu$$

5.4.2 Lemma: $g^{\theta_0, \alpha}$ is a metric in some neighbourhood of θ_0 for all $\alpha \in \mathbb{R}$

Proof As in the previous lemma $g^{\theta_0, \alpha}$ transforms as a tensor. Locally near θ_0 it will be non-singular since, at the preferred point, $g^{\theta_0, \alpha}$ equals the Fisher information.

Q.E.D.

Thus $g^{\theta_0, \alpha}$ is another preferred point metric which we shall view as a perturbation of the Fisher metric. As before we shall expand $p(x, \theta_0)$ as a Taylor series to produce a power series approximation of the Fisher information.

5.4.3 Theorem: To first order the Levi-Civita connection of $g^{\theta_0, \alpha}$ is the $(\alpha-1)$ -connection.

Proof. Recall that

$$g_{ij}^{\theta_0, \alpha} = \int_{\mathbf{x}} \left(\frac{p(\mathbf{x}, \theta)}{p(\mathbf{x}, \theta_0)} \right)^\alpha \frac{\partial}{\partial \theta_i} \ln p(\mathbf{x}, \theta) \cdot \frac{\partial}{\partial \theta_j} \ln p(\mathbf{x}, \theta) \cdot p(\mathbf{x}, \theta_0) \cdot d\mu \quad (\dagger)$$

Therefore using

$$p(\mathbf{x}, \theta_0) = p(\mathbf{x}, \theta) - (\theta - \theta_0)^k \frac{\partial}{\partial \theta_k} p(\mathbf{x}, \theta) + O((\theta - \theta_0)^2)$$

(\dagger) expands to:

$$\begin{aligned} & \int_{\mathbf{x}} \left(\frac{p(\mathbf{x}, \theta)}{p(\mathbf{x}, \theta) - (\theta - \theta_0)^k \frac{\partial}{\partial \theta_k} p(\mathbf{x}, \theta) + O((\theta - \theta_0)^2)} \right)^\alpha \frac{\partial}{\partial \theta_i} \ln p(\mathbf{x}, \theta) \cdot \frac{\partial}{\partial \theta_j} \ln p(\mathbf{x}, \theta) \times \\ & \quad (p(\mathbf{x}, \theta) - (\theta - \theta_0)^k \frac{\partial}{\partial \theta_k} p(\mathbf{x}, \theta) + O((\theta - \theta_0)^2)) d\mu \\ &= \int_{\mathbf{x}} \partial_i \ln p(\mathbf{x}, \theta) \partial_j \ln p(\mathbf{x}, \theta) \cdot \left(1 + \frac{\alpha}{p(\mathbf{x}, \theta)} (\theta - \theta_0)^k \partial_k p(\mathbf{x}, \theta) + O((\theta - \theta_0)^2) \right) \times \\ & \quad (p(\mathbf{x}, \theta) - (\theta - \theta_0)^k \frac{\partial}{\partial \theta_k} p(\mathbf{x}, \theta) + O((\theta - \theta_0)^2)) d\mu \\ &= \int_{\mathbf{x}} \partial_i \ln p(\mathbf{x}, \theta) \partial_j \ln p(\mathbf{x}, \theta) p(\mathbf{x}, \theta) d\mu + (\theta - \theta_0)^k \int_{\mathbf{x}} \frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j} (\alpha - 1) \frac{\partial l}{\partial \theta_k} p(\mathbf{x}, \theta) d\mu + \\ & \quad O((\theta - \theta_0)^2) \end{aligned}$$

Preferred Point Geometry

$$= g_{ij} + (\theta - \theta_0)^k \left((\alpha - 1) \int \frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j} \frac{\partial l}{\partial \theta_k} p(x, \theta) d\mu \right) + O((\theta - \theta_0)^2)$$

So, again, this is a power series expansion of the form of Theorem 5.3.2. Hence the Levi-Civita connection of the perturbed metric is

$$\Gamma_{ijk} + (\alpha - 1)/2 T_{ijk} + O(\theta - \theta_0)$$

where Γ_{ijk} are the Christoffel symbols for the Fisher information metric and T_{ijk} is Lauritzen's skewness tensor.

Q.E.D.

Hence we have shown that there exists a one parameter family of preferred point metrics each of which has a connection which equals one of the α -connections at θ_0 . In general the α -connection is a first order approximation of one of these preferred point metrics. Thus the theory of Statistical manifolds can be looked at as a first order approximation to a more general theory of preferred point geometry. To see if this is a useful generalisation or not we shall, in the rest of the chapter, study various aspects of Statistical manifold theory in the larger context of preferred point geometry.

5.5 Applications to Asymptotic Analysis.

One of the main criteria for whether we have a good generalisation will be the extent to which the preferred point view answers the questions which we posed at the beginning of this chapter. Specifically, will the preferred point geometry be the unifying principle which connects the three different applications of the α -connections? In this section we will study its application to higher order asymptotic analysis. We shall show that this does fit naturally into a preferred point context. As before we shall be able to identify the skewness tensor as the first order correction to the curvature.

We shall review the justification for using the Fisher information first. The reason comes from looking at the asymptotic expansion of the maximum likelihood estimate as the sample size, n , increases. The first term of the asymptotic expansion shows that to order $(1/\sqrt{n})$ the distribution is normal with a variance

which is the inverse of the Fisher information at θ_0 , (see Chapter 2). The theorems of the previous section tell us that in calculating the curvature to a particular order, say $(1/\sqrt{n})$, we have to be careful of higher order terms in the expansion of the metric. We can see that at θ_0 we would like to have the Fisher information as the metric but elsewhere what the metric should be is less clear.

We shall consider the role of the metric of Example 5.4.

$$g^{\theta_0}_{ij} = E_{p(x,\theta_0)}[\partial_i \ln p(x,\theta) . \partial_j \ln p(x,\theta)]$$

$$= E_{\theta_0}[\partial_i \ln p(x,\theta) . \partial_j \ln p(x,\theta)], \text{ say.}$$

For the moment we shall not look at any theoretical justification for using this apart from noting that at the true parameter g^{θ_0} is equal to the Fisher information. Thus the justification for using the Fisher information equally applies to g^{θ_0} . We shall see that working in the preferred point system brings a considerable clarity to some of Amari's results. We shall not, in this section, prove anything new. Rather we shall just reinterpret some of Amari's results in a preferred point context. We shall show how doing this gives much more geometric insight to the theorems.

We shall work in the same framework as [Amari], and to aid comparison of the results we shall use his notation.

5.5.1 Definition A family $q(x,u)$ of distributions parametrised by a vector parameter u is said to be a curved exponential family when the density function is

$$q(x,u) = \exp\{ \theta^i(u)x_i - \psi(\theta(u)) \}$$

($\theta = \theta(u)$ is a (vector-valued) function of u).

We let $u = (u^\alpha)$ ($\alpha = 1, 2, \dots, m$) be an m -dimensional parameter, while $\theta = (\theta^i)$ ($i = 1, 2, \dots, n$) is an n -dimensional one and $m < n$. In this case M is an (n, m) -curved exponential family.

5.5.2 Definition: M is a curved exponential family embedded in some full exponential family, S . To each point $u \in M$ we attach an $(n-m)$ -dimensional submanifold $A(u)$ in such a way that $A(u)$ is a local foliation of S , which is parametrised by M . The set $A(u)$ is called the ancillary family.

$A(u)$ is a local foliation so we can use it to define a local coordinate system in S . We let $v = (v^k)$ ($k = m+1, \dots, n$) be a coordinate system in $A(u)$. The

Preferred Point Geometry

origin $v = 0$ is put at $\theta(u)$ on M . Then we choose v for each $A(u)$ such that v varies smoothly enough such that the pair

$$w = (u, v)$$

is a local coordinate system in S .

We shall use the indices a, b, c, \dots running from 1 to m to denote quantities related to M , and the indices $\kappa, \lambda, \mu, \dots$ running from $m+1$ to n to those relating to $A(u)$.

We shall now consider the inference problem for a point in $M \subset S$. Let $\hat{\theta}$ denote the maximum likelihood estimator in S . We shall consider how the value in S should be used to draw inference on a point in M . We can consider our estimation procedure as being a function from S to M . Thus any procedure u^* can be defined by the inverse image of its corresponding function

$$A(u) = \{\eta \in S \mid u = u^*(\eta)\}$$

i.e. to each estimate we can associate an ancillary family. Amari then uses the geometry of this ancillary family to study the efficiency of the estimator. We shall quote his relevant theorems.

5.5.3 Theorem (Amari): An estimator u^* is consistent when and only when every point $\eta(u) \in M \subset S$ is included in the associated ancillary family.

Proof. Page 129 [Amari].

5.5.4. Theorem (Amari): A consistent estimator u^* is first order efficient iff its associated ancillary family meets M orthogonally in the Fisher metric.

Proof. page 131 [Amari].

Amari then shows that unbiased first order efficient estimators are always second order efficient. Finally he proves the following.

5.5.5 Theorem (Amari): An estimate is third order efficient when and only when the associated ancillary family $A(u)$ has zero -1-curvature at $v=0$

Proof. Page 134

Preferred Point Geometry

Thus in going from first to third order efficiency Amari uses first the Fisher information (0-connection), and then the -1-connection. Thus here is an example of the mixing of two different types of geometric structure which we spoke about at the beginning of this chapter. We shall see how we can interpret 5.5.4 and 5.5.5 in one preferred point structure. The structure we shall use is the $E_{\theta_0}[\partial_i, \partial_j]$ geometry where θ_0 is the true parameter.

5.5.6 Lemma: If $A(u)$ is orthogonal to M at θ_0 in the Fisher metric then it is orthogonal to M at θ_0 in the $E_{\theta_0}[\partial_i, \partial_j]$ metric.

Proof. At θ_0 the two metrics agree.

5.5.7 Lemma: At θ_0 the Levi-Civita curvature of the metric agrees exactly with the -1-connection.

Proof. We have seen that they agree to first order in $(\theta - \theta_0)$ thus at θ_0 they agree exactly.

Thus in view of the above two Lemmas we can rewrite Theorems 5.5.4 and 5.5.5 as the more natural geometric theorem.

5.5.8 Theorem: Let us use the $E_{\theta_0}[\partial_i, \partial_j]$ geometry. An (unbiased) estimator is first (second) order efficient if $A(u)$ cuts M at θ_0 orthogonally and is third order efficient if $A(u)$ also has zero curvature.

The result of this theorem can be interpreted geometrically in the following way. We are looking at estimators on the curved exponential family N as being projections from the full exponential family M in which N lies. If we are working in the $E_{\theta_0}[\partial_i, \partial_j]$ geometry then the natural projection is where each point in M goes to the point $\theta_0 \in N$ which is closest in the $E_{\theta_0}[\partial_i, \partial_j]$ geometry. This means that the estimate is mapped to $\theta_0 \in N$ via the $E_{\theta_0}[\partial_i, \partial_j]$ geodesic which cuts N orthogonally. Since all geodesics have zero curvature everywhere this one has zero curvature at θ_0 . Hence this projection is third order efficient. Thus we can see that in the preferred point geometry the most natural geometric projection is also the best statistically.

We leave as an open problem the question whether this natural geometric estimator is in fact better than third order efficient.

5.6 Divergence Measures and Preferred Point Geometries.

In this section we shall consider the relationship between general divergence functions and divergence functions which are derived from geodesic distances. In this second group we shall compare two types. The first is based on the geodesic distance for a fixed Riemannian geometry. In the second group we shall look at the geodesics of a preferred point geometry.

By the term divergence function we shall mean any function

$$d:M \times M \rightarrow \mathbb{R}$$

which acts as a distance measure. We shall require that d be smooth and we shall assume that the derivative of d has full rank apart from at the points (θ, θ) . Because d has an interpretation as a distance we shall also require that d satisfies the condition that,

$$d(x, y) \geq 0 \text{ for all } (x, y) \in M^2 \text{ and that } d(x, y) = 0 \text{ iff } x = y.$$

Note that we will not have, in general, a triangle inequality.

Many divergence functions that we shall study will be quadratic in the following sense. For a fixed point x and a path $y(t)$ which goes through x , at $t=0$ say, then

$$\left. \frac{d}{dt} d(x, y(t)) \right|_{t=0} = 0$$

This is a very common condition on many divergence functions which are used in statistics. Clearly for any divergence function d we can find a quadratic divergence function which will be equivalent by considering d^2 .

In [Rao] it is shown that at a point θ_0 certain well known quadratic divergence functions can be linearised and the result is a metric in the tangent space at θ_0 . In the following sections we shall see how this type of result can be generalised to find a metric at every point of the manifold. Thus we shall bring the tools of differential geometry to the study of divergence functions.

Let us consider first the relationship between Riemannian geometry and divergence functions. We have the natural question; given a general divergence function, $d(,)$, is there a metric whose geodesic distances agree with those of d ? In general the answer to this question is no. If we have no restriction on $d(,)$ then

there is no reason to think that it is a symmetric function, whereas geodesic distances always are i.e. it will not be true in general that

$$d(P,Q) = d(Q,P).$$

This non-symmetry is a reasonable property to consider. One of the most widely used divergence functions, the Kullback-Leibler divergence, is obviously non-symmetric.

It is the fact that the geometry of statistical spaces does not possess a natural symmetry that gives rise to the need for using preferred point geometry if we want to use differential geometric methods. The lack of symmetry is a very deep property of statistical modeling as we can see if we consider, for example, the Neyman-Pearson Lemma, (see [Cox and Hinkley]). There the lack of symmetry is expressed in the fundamental difference between *size* and *power*. It is the assumption of which element of the parametric family is the true distribution that is critical in the analysis behind this lemma. Our new geometric structure also reflects this fact whereas Riemannian geometry does not.

If, however, we only restrict our attention to symmetric divergences we still see that there won't be a one to one correspondence between divergence functions and geodesic distances. The following arguments demonstrate this.

Firstly, we can see that for a Riemannian based geometry the triangle inequality always holds for geodesic distances, at least in a small enough neighbourhood. This is certainly not true of general divergence functions. Also the two types of function are fundamentally different in the following sense. Any geodesic distance is essentially a finite dimensional object. Once p points have been fixed on a p -dimensional manifold then the metric at any other point is determined by the p geodesic arcs joining it to the others. This is not going to be true of a general function $d(,)$ which should be seen as a family of real valued functions on the manifold which has been parametrised by the manifold itself. Thus it is essentially an infinite dimensional object.

We can therefore conclude, from all these arguments, that there is no one to one correspondence between divergence functions (even symmetric ones) and Riemannian geometries.

Now let us consider the relationship between divergence functions and preferred point geometries. We shall pick θ_0 as our preferred point, that is θ_0 corresponds to the true distribution. Further, let us suppose that $d(,)$ is our divergence function. We shall consider the following:

Preferred Point Geometry

$$d_{\theta_0}(\theta) = d(\theta_0, \theta)$$

We shall call this restricted function the *preferred point divergence*, (or just divergence if the preferred point is clearly understood).

Note also that we could have defined

$$d'_{\theta_0}(\theta) = d(\theta, \theta_0).$$

All the following theory will work with such a definition. Often with statistical examples of divergence functions there will be a natural choice of which of the two forms is preferable. For example with the Kullback-Liebler divergence we would want to take expectation with respect to the true distribution.

The function d_{θ_0} is completely determined by the level sets

$$S_{d_{\theta_0}}(c) = \{\theta \mid d_{\theta_0}(\theta) = c\}.$$

If $d_{\theta_0}(\theta)$ is clearly understood we shall define

$$S(c) = S_{d_{\theta_0}}(c).$$

We call these sets the *divergence c-spheres* based at θ_0 .

5.6.1. Note: There is an equivalence relation on the set of preferred point divergence based at θ_0 . It is defined by:

$$d_{\theta_0}(\theta) \sim d'_{\theta_0}(\theta) \quad \text{iff} \quad \text{for all } c > 0 \text{ there exists } c' > 0 \text{ such that } S_{d_{\theta_0}}(c) = S_{d'_{\theta_0}}(c').$$

That is to say that two divergences are equivalent if their level sets agree, as sets, but not necessarily on the value which is put on each set. This equivalence relation is clearly important but we do not follow up this idea here.

We shall now show that given any preferred point divergence we can find a preferred point metric which is compatible for all distances measured from the preferred point. This is equivalent to showing that there exists a complete preferred point geometry which is equivalent to any divergence function. Note that all geodesic distances are measured from the relevant preferred point, i.e. if we talk about the geodesic distance from θ_1 to θ_2 we measure it in the θ_1 -geometry.

Preferred Point Geometry

We shall in fact show that there are many such preferred point metrics and we shall discuss what extra information is contained in a metric system over a divergence function.

We start by describing a construction which enables us to generate a preferred point geometry from a divergence function. This is not a canonical construction and contains a number of choices. We shall explore the statistical significance of these choices later in this chapter.

5.6.2. Theorem: Let $d_{\theta_0}(\theta)$ be a divergence function, then there exists a preferred point metric $g_{\theta_0}(\cdot, \cdot)$ which is compatible with $d_{\theta_0}(\theta)$, i.e. if $d_g(\theta)$ is the θ_0 -geodesic distance of θ_0 to θ , then

$$d_{\theta_0}(\theta) = d_g(\theta)$$

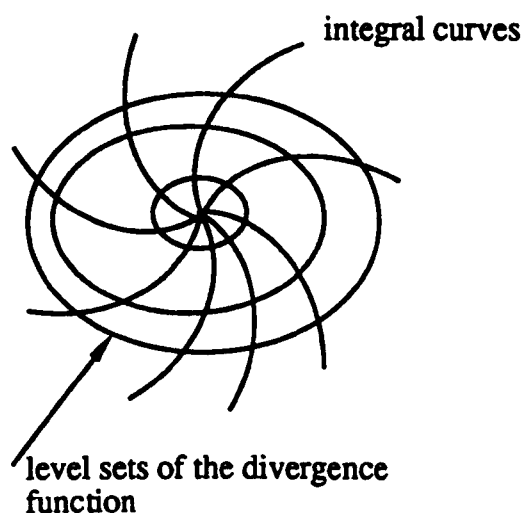
for all θ in some neighbourhood of θ_0 .

Proof: First we choose a vector field $X(\theta)$ which has a singularity at θ_0 and is otherwise nonsingular in some neighbourhood of θ_0 . We choose X such that DX is a +ve definite matrix. Such a singularity is called a source. This field can be chosen arbitrarily with the restriction that it must be nowhere tangential to the level sets $S_d(c)$. For a reference to basis vector field theory on a manifold see [Palis & deMelo].

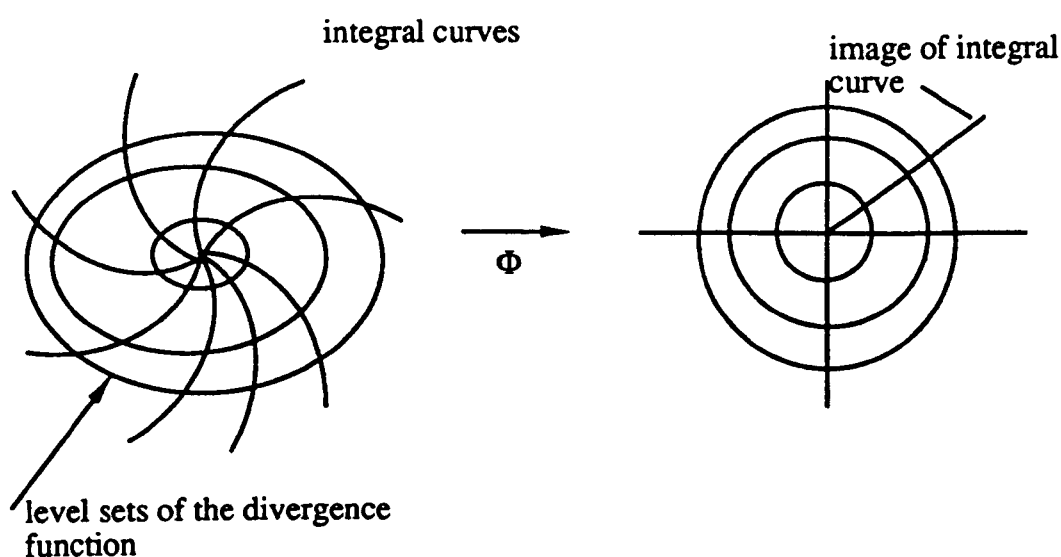
By the usual existence and uniqueness theorems for the existence of flows along vector fields (see Abraham, Marsden and Ratiu, [A,M&R]) we know that we can locally find a set of integral curves to the vector field which pass through θ_0 .

Thus we can put on our manifold the 'polar coordinates' defined by the flow lines of $X(\theta)$ and the level sets of d_{θ_0} . In a neighbourhood of θ_0 this will be a proper coordinate system for a neighbourhood of M around θ_0 , as the following argument shows.

Preferred Point Geometry



We define a map Φ from our manifold M to \mathbb{R}^p , where p is the dimension of M . This map will only be defined in the neighbourhood of θ_0 where we have the 'polar coordinates' defined.



We first map θ_0 to the origin of \mathbb{R}^p . We then choose a basis of \mathbb{R}^p and a basis for TM_{θ_0} , the tangent space to M at θ_0 . We write down the isomorphism ϕ between these two vector spaces which maps one basis to the other. To extend the map away from θ_0 you pick a tangent vector $v \in TM_{\theta_0}$. This is mapped to $\phi(v) \in \mathbb{R}^p$. Then we map the flow line in M which starts at θ_0 with initial direction

Preferred Point Geometry

v to the line in \mathbb{R}^p spanned by $\phi(v)$. This will be well defined by the uniqueness theorem for flows.

We can then completely define Φ by letting the level set $S(c)$ map to the Euclidean sphere of radius c centred on the origin. Thus all we have done is to map the nonlinear 'polar coordinates' on M to the standard polar coordinates on \mathbb{R}^p . Since the vector field X is never tangential to $S(c)$ we see that the derivative of Φ is always of full rank. Hence by the inverse function theorem Φ will be a diffeomorphism and so will define a genuine coordinate system.

We now define a metric on M by pulling back the standard metric on \mathbb{R}^p via the map Φ . Thus we define $g(,)$ by

$$g_{\theta_0}(v_1, v_2) = \langle \Phi^*v_1, \Phi^*v_2 \rangle$$

where \langle, \rangle is the standard Euclidean inner product on \mathbb{R}^p , and $\Phi^*: TM \rightarrow \mathbb{R}^p$ is the lift of Φ to the relevant tangent spaces.

We shall show that the g_{θ_0} -geodesic distance from θ_0 is equal to the divergence from θ_0 . This is straightforward since by construction g_{θ_0} is isometric to the standard metric on \mathbb{R}^p via Φ . Hence Φ maps g_{θ_0} -geodesic spheres centred at θ_0 to Euclidean spheres centred at the origin of \mathbb{R}^p . However, we know that Φ maps the level sets of d_{θ_0} to these same Euclidean spheres. Thus we conclude that $S_{d_{\theta_0}}(c)$ must be a g_{θ_0} -geodesic sphere of the correct radius. This completes the proof.

Q.E.D.

5.6.3. Corollary: For any divergence function d_{θ_0} there is a flat preferred point metric compatible with d_{θ_0} , i.e. the metric has zero Riemann-Christoffel curvature for each θ_0 .

Proof: We have constructed this metric in the previous theorem.

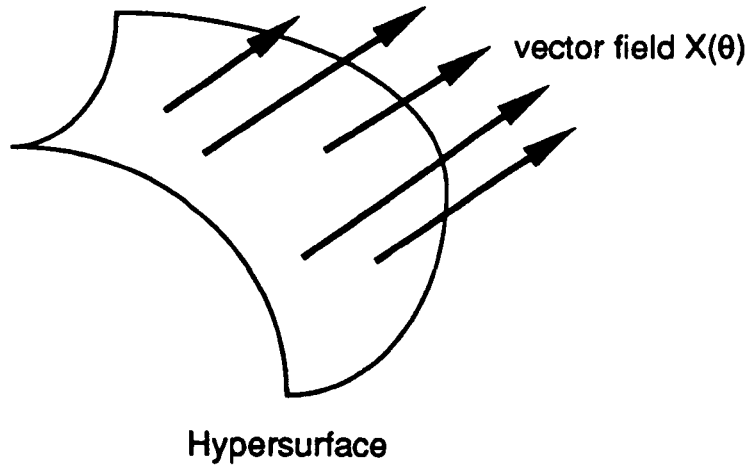
Q.E.D.

We can see therefore that the theory of divergence functions fits very naturally into a preferred point geometry context. The correspondence is far from one to one however. The proof of Theorem 5.6.2 can be changed to produce a metric of any curvature we want. Also there is still room for choice about the vector field used in the construction. We would therefore like to know what extra information

is contained in a preferred point geometry in comparison to one based on a divergence function. We would also like to know if this information can be fixed by statistical considerations.

5.7. Choice of vector field.

First we shall discuss the choice of the vector field $X(\theta)$ which is used in our construction of a compatible metric in Theorem 5.6.2. We recall that there are two conditions needed for $X(\theta)$ to be used. The first is that $X(\theta)$ is non-singular in some open neighbourhood of θ_0 , apart from at θ_0 where the singularity is a source, see [Palis&deMelo]. The second condition is that the vector field is always transverse to the level sets of d_{θ_0} . This second condition can be expressed as the fact that $X(\theta)$ does not lie in the tangent space to $S_{d_{\theta_0}}(d_{\theta_0}(\theta))$. So at each point θ ($\neq \theta_0$) by the nondegeneracy of d_{θ_0} we have a hypersurface $S_{d_{\theta_0}}(d_{\theta_0}(\theta))$ and a vector $X(\theta)$ which does not lie in its tangent space. See figure below.



We can now see what the vector field does in our construction of the metric. At each point θ it defines a direction which will be the orthogonal direction to $S_{d_{\theta_0}}(d_{\theta_0}(\theta))$. We shall show this in the following theorem.

5.7.1. Theorem: Let $X_1(\theta)$ and $X_2(\theta)$ be two vector fields with the following relationship

$$X_1(\theta) = f(\theta).X_2(\theta)$$

where f is a positive real valued smooth function on M .

Replacing $X_1(\theta)$ with $X_2(\theta)$ in the construction of Theorem 5.6.2 does not change the metric constructed.

Proof: We shall show that the choice of metric depends only on the flow lines of the vector field used. We shall therefore prove the result by demonstrating that the two vector fields give the same flow lines.

The construction of the metric is only dependent on the construction of the map $\Phi:M \rightarrow \mathbb{R}^p$, and the construction of this map followed the formula;

- (i) Define the isomorphism from the tangent space at θ_0 to $0 \in \mathbb{R}^p$.
- (ii) Map the flow lines of the vector field to lines through the origin.
- (iii) Define Φ on each flow line by letting d_{θ_0} correspond to the standard

Euclidean distance on \mathbb{R}^p .

Thus the vector fields are only used in the construction via the flow lines.

We see that $X_1(\theta)$ and $X_2(\theta)$ have the same flow lines as they are related by a simple reparametrisation.

Q.E.D.

This result can be interpreted as saying that it is the choice of direction of the vector field which is relevant to the construction of the metric. We note from the proof of the theorem that we have the following

5.7.2. Corollary: If two vector fields generate the same flow lines then they generate the same metric.

5.7.3. Lemma: The flow lines of the vector field used in the construction are the geodesic lines from θ_0 in the induced g_{θ_0} -metric.

Proof. By a well known result of the calculus of variations or by taking geodesic normal coordinates (see for example [Auslander]) we see that the geodesics

starting at θ_0 are all orthogonal to the geodesic spheres centred at θ_0 . By construction, the geodesic spheres are the level sets of d_{θ_0} . Further since it is true that lines through the origin of \mathbb{R}^p are orthogonal to the Euclidean spheres centred at the origin then our construction ensures that $X(\theta)$ is orthogonal to these level sets of d_{θ_0} . Hence we see that since the level set is a hypersurface we must have that $X(\theta)$ is parallel to the geodesic through θ . Thus the flow lines of $X(\theta)$ are geodesics.

Q.E.D.

This observation gives an alternative view of the choice of vector field. Rather than viewing the choice as one of direction we can view it as a choice of which lines are to be the geodesics from θ_0 in our g_{θ_0} -metric.

From the above result we can see that we can further extend the result of Corollary 5.6.3. We shall show that not only can we pick a compatible metric which is flat, i.e. one which has a set of affine coordinates, we can in fact choose, under certain regularity conditions, which coordinates we would like to be affine.

5.7.4. Theorem: For any divergence function $d_{\theta_0}(\cdot)$ and any non-singular coordinate system θ , there exists a flat metric compatible with d_{θ_0} and whose geodesics from θ_0 are given by the lines $\{\theta_0 + \lambda(\theta - \theta_0) | \theta \in M\}$ given that these lines are always transverse to the level sets of d_{θ_0} .

Proof. We can see that the tangent field to the lines $\{\theta_0 + \lambda(\theta - \theta_0) | \theta \in M\}$ satisfy the conditions of the construction. Thus if we use this vector field we have the result by Lemma 5.7.3.

Q.E.D.

We shall now look briefly at the statistical implications of these considerations. Since there are two ways of looking at the choice of the vector field we shall give two possible statistical 'natural' choices.

We can view the choice of vector field as a choice of which lines we want to be the geodesics from θ_0 . If we did this and there were a natural statistical set of coordinates then these could be used as our choice of affine coordinates. For example in the exponential family $\{\exp(\theta^i x_i + \phi(\theta))\}$ the coordinates θ could be used.

Preferred Point Geometry

If however the choice of vector field is to be decided by which direction we wish to consider orthogonal to the level sets of d_{θ_0} then we can take the following definition. This definition is however purely *ad hoc*. There are clearly many possible definitions of orthogonality. We include this one as an example.

Take the preferred point viewpoint and let θ_0 represent the true parameter. Suppose that T_1, \dots, T_{p-1} span the tangent space of the level set of d_{θ_0} at θ . Then we can define the vector field $X(\theta)$ as the solution to the following equations:

$$E_{\theta_0}[T_i X(\theta)] = 0 \quad \text{for all } i=1, \dots, p-1.$$

5.8 Choice of curvature.

We still have a lot of choice over our construction of a compatible metric even if we have decided on the choice of vector field. For instance it is quite easy to see that the choice of curvature of our preferred point metric is still completely unconstrained. We can interpret the results of 5.6.2 and 5.6.3 as telling us this. It is therefore natural to ask what extra information do we have to add to a divergence function to make a unique choice of compatible (preferred point) metric? We shall proceed by trying to answer this question in the abstract and then trying to apply statistical considerations to fix which is the best choice for our requirements.

The construction of our metric relies on producing a map Φ from our manifold to some target Riemannian manifold T , say. We note that in any actual example you would have to know a closed form formula for the geodesic distance on T . This is why R^p is an ideal choice. However the proof of Theorem 5.6.2 would go through with the natural modifications for any target Riemannian manifold. It is by choosing this manifold that we achieve the choice of curvature for the metric. The construction forces the metric to have the curvature of the target manifold.

We shall look at what extra information is carried by a metric compared to a divergence function when they are compatible. We know that they agree on lengths as measured from θ_0 . However a metric can measure more than just the

Preferred Point Geometry

lengths of paths. Given a metric on a manifold there is an induced p -dimensional volume measure. In $(\theta^1, \dots, \theta^p)$ coordinates this is given by

$$\det (g_{ij})d\theta^1\dots d\theta^p \quad (*)$$

if g_{ij} is the metric.

Furthermore, if N is an n -dimensional submanifold of M , then the metric on M induces a metric on N . Thus by applying $(*)$ to N we see that the metric g_{ij} in fact induces an n -dimensional volume form on N .

Thus a metric tells you how to measure volumes on the manifold M and on all its submanifolds.

We shall now study the situation on a two dimensional manifold.

5.8.1. Theorem: Let S be a two dimensional manifold and $d_{\theta_0}(\theta)$ a divergence function on S . Let $\mu(\theta)$ be an area measure on S . We select our vector field $X(\theta)$, and choice of metric at θ_0 , $g_{\theta_0}(\theta_0)$. Then there exists, locally, a unique metric $g_{\theta_0}(\theta)$ compatible with the above in the sense that

- (I) $g_{\theta_0}(\theta)$ agrees with the choice of metric at θ_0 .
- (II) $X(\theta)$ is orthogonal to the level sets of d_{θ_0} when we use the $g_{\theta_0}(\theta)$ metric.
- (III) The geodesic distance from θ_0 to θ agrees with the divergence.
- (IV) The area measure induced by $g_{\theta_0}(\theta)$ agrees with $\mu(\theta)$.

Proof. We shall use the coordinate system defined by the level sets of d_{θ_0} and the vector field $X(\theta)$. I.e. at $\theta(\neq\theta_0)$ we shall define the basis of the tangent space to be $\{\partial_1, X(\theta)\}$. Where ∂_1 is the tangent to the level set $S_{d_{\theta_0}}(d_{\theta_0}(\theta))$. By assumption ∂_1 and $X(\theta)$ are linearly independent.

By condition (II) we can see that with respect to this coordinate system the metric is of the form:

$$\begin{bmatrix} g_{11} & 0 \\ 0 & g_{22} \end{bmatrix}$$

Preferred Point Geometry

Consider the geodesic normal coordinates for a metric around the point θ_0 . These are defined by the geodesics from θ_0 and the geodesic spheres centred at θ_0 . By a well known theorem (see for example page 179 [Auslander]) we see that in these coordinates for g_{θ_0} the metric has the form:

$$\begin{bmatrix} g'_{11} & 0 \\ 0 & g'_{22} \end{bmatrix}$$

By (III) we see that the geodesic balls centred at θ_0 agree with the level sets of the function d_{θ_0} . We can deduce that $g_{11}(\theta) = g'_{11}(\theta)$ for all $\theta \neq \theta_0$. The reason for this is that the normals to the level sets of d_{θ_0} are parallel to the tangents to the geodesics from θ_0 . Hence the vector field $X(\theta)$ defines the geodesics because they are its integral curves. Therefore to find the geodesic distance we must integrate along the level curves of X . Thus we get the geodesic distance as

$$\int_0^{d(\theta)} \sqrt{g_{11}}$$

Therefore condition (III) determines g_{11} .

We have not yet fixed g_{22} however and it is this which determines the curvature of the metric. To do this we use condition (IV). We have that

$$\det g_{ij} = g_{11}g_{22} = \mu(\theta).$$

This equation is enough to fix the metric uniquely

Q.E.D.

Thus in the two dimensional case we must add an area function to the divergence to fix the metric uniquely. If we wish to apply these ideas to statistics it means we must define a statistically sensible area measure to our parameter spaces.

In the general p -dimensional case the information needed to specify the metric is similar. We can see that we would need to define a p -dimensional measure as well as a measure on all the submanifolds in a consistent way. This is however not easy to do without using a metric type notation. Thus the p -

dimensional equivalent to Theorem 5.8.1 is not nearly so useful. However we do have the following result.

5.8.2 Theorem: Given the conditions of 5.8.1 except we now work in p -dimensions. Consider the coordinates defined by the vector field $X(\theta)$ and the coordinates on the level sets of d_{θ_0} . We shall define a $p-1$ dimensional metric g^t on the level sets of d_{θ_0} . Then there is a unique metric which satisfies (I),(II,) and(III) of theorem 5.8.1 and also

(IV) g_{θ_0} agrees with g^t when restricted to the level sets of d_{θ_0} .

Proof. The proof is simply that with respect to the coordinates we have taken the metric d_{θ_0} must be of the form

$$\left[\begin{array}{c|c} g_{11} & 0 \\ \hline 0 & g_{ij}^* \end{array} \right]$$

Then condition (III) fixes g_{11} and (IV) fixes $g_{ij}^* = g_{ij}^t$.

Q.E.D.

Thus we have the viewpoint that the essential information needed to fix a preferred point metric to make it compatible with a divergence function is a set of metrics on the level sets of d_{θ_0} . We can see that these metrics will fix any area or n -dimensional volume measures in the manifold.

We have therefore shown how the two structures of divergence functions and preferred point geometries interact in general. In [Amari] there are connected results on the relationship between the α -connections structure and some families of divergence functions. These results hold in what he defines as α -flat families of distributions of which the exponential families are examples. It would be interesting to find the connections between his work and the results of the previous few sections. We refer forward to Chapter 6 for a powerful result in this direction. Amari's theorems produce results connecting the Kullback-Leibler divergence with the -1 -connection. In Chapter 6 we produce the preferred point metric which corresponds to the -1 -connection and then in the exponential family case we show the exact geometric relationship between the Kullback-Leibler divergence and the geodesic distance in this preferred point geometry. There would seem to be considerable scope for further results along these lines.

5.9 Conclusion.

At the start of this chapter we asked some questions about the geometric and statistical naturalness of the Statistical manifold structure. In trying to produce a more natural one we have defined a new geometric structure which we feel is closer to the natural framework in which the geometrisation of statistics can take place. This is true both from a geometric and from a statistical point of view.

Following this idea we have produced results in two main areas in which we show the power of the preferred point idea in applications. These areas are divergence function theory and asymptotic analysis.

It must be said, however, that this analysis is very incomplete, and much further work remains to be done. In particular we have not looked at the area of reparametrisation in which the Statistical manifold theory has important results. Also there are aspects of Statistical manifold theory which have not yet been understood in the preferred point context. One very important example of this is the powerful ideas of duality which exist in Statistical manifold theory. We feel that there are parallels to this idea in preferred point geometry. For instance if we have two points, one the true parameter and the other some estimate, then there is a form of duality if the roles of these two point are exchanged. It would be interesting to develop this idea and see if we can use it to find the statistical significance of the geometric theorems of Lauritzen concerning duality.

Chapter Six

Examples of Preferred Point Geometries in Statistics.

6.1 Introduction.

This chapter is concerned with looking at explicit statistical examples of preferred point geometries. We have already seen a couple of these. We shall, in each case, define a particular metric and discuss the statistical motivation for its use. To get a set of explicit examples we shall apply each metric to an exponential family. Often the simplicity of the representation of the exponential family will allow us to gain considerable statistical and geometric information.

We will bring out in each example the relationship, which we have already, seen between the preferred point geometry and Amari's results using a Statistical manifold structure.

We shall also consider the particular nature of the metric, from its particular preferred point nature. These observations will be the start of some basic preferred point geometry theory which we shall look at in Chapter 7.

6.2 The Embedding metric g^{θ_0} .

The first example is the natural preferred point extension of the Fisher information $I(\theta)$. We recall that we defined the Fisher information as,

$$I(\theta) = E_{\theta} [\partial_i \ln p(x, \theta) . \partial_j \ln p(x, \theta)]$$

where $\partial_i = \partial / \partial \theta_i$.

The preferred point viewpoint is that one distribution, θ_0 , is on a different footing than the others. Hence if we are viewing θ_0 as the true distribution then it is natural to consider

Examples.

$$g^{\theta_0} = E_{\theta_0} [\partial_i \ln p(x, \theta) \cdot \partial_j \ln p(x, \theta)]$$

as a candidate for a preferred point metric. We shall denote this quantity by g^{θ_0} and call it the *preferred point expected information* metric.

6.2.1 Lemma: g^{θ_0} a metric in a neighbourhood of θ_0 .

Proof. See Lemma 5.4.1

The local restriction on g^{θ_0} should be noted. It is a common property of preferred point metrics. It would be interesting to study the set of singular points for g^{θ_0} . This set will of course vary for each θ_0 . However we will not develop this idea any further in this thesis.

We cannot yet produce any direct statistical arguments which are fully convincing for the introduction of any of our preferred point metrics. We refer forward to Chapter 8 for some possible lines of approach. For the moment we shall defend their use by two types of argument. We shall present some heuristic justification firstly, and secondly the usefulness of the results which the metrics generate will act as their justification.

We start with a heuristic argument. A metric is a quadratic form on each tangent space of the manifold. Thus we can follow Amari in considering the tangent spaces to be spanned by $\{\partial_i \ln p\}_{i=1, \dots, p}$. $E_{\theta_0}[\]$ is a natural quadratic form on this tangent space. It is telling us the second moment matrix of the vectors in the tangent space which would be generated at θ when θ_0 is the true distribution. In this viewpoint it is considerably more natural to use $E_{\theta_0}[\]$ rather than $E_{\theta}[\]$ to generate the matrix.

We have, in Chapter 5, already noted that the first order approximation to the Levi-Civita connection of g^{θ_0} is given by the -1 -connection of Amari/Lauritzen. Hence the preferred point metric inherits the results on asymptotic theory which use the -1 -connection.

We shall now go on to calculate the precise form of the connection.

Examples.

6.2.2 Lemma: If we let $\Gamma_{ijk}^{\theta_0}$ denote the Christoffel symbols for the Levi-Civita connection of g^{θ_0} , then

$$\Gamma_{ijk}^{\theta_0} = E_{\theta_0} [\partial_{ij}^2 l \cdot \partial_k l]$$

where $\partial_{ij}^2 l = \partial^2 / \partial \theta_i \partial \theta_j (\ln p(x, \theta))$.

Proof. This is a simple calculation using the formula that defines the Christoffel symbols for a metric. Thus

$$\Gamma_{ijk}^{\theta_0} = \frac{1}{2} \left(\frac{\partial g_{kj}^{\theta_0}}{\partial \theta_i} + \frac{\partial g_{ki}^{\theta_0}}{\partial \theta_j} - \frac{\partial g_{ij}^{\theta_0}}{\partial \theta_k} \right) \quad (*)$$

Since

$$\begin{aligned} g_{ij}^{\theta_0} &= E[\partial_i l \partial_j l] \\ &= \int_X \partial_i l(\theta) \cdot \partial_j l(\theta) \cdot p(x, \theta_0) dx \end{aligned}$$

By differentiating through the integral sign which the regularity conditions of Chapter 1 allows us to do, we get

$$\frac{\partial}{\partial \theta_k} g_{ij}^{\theta_0} = \int_X [\partial_{ik}^2 l(\theta) \cdot \partial_j l(\theta) + \partial_i l(\theta) \cdot \partial_{jk}^2 l(\theta)] p(x, \theta_0) dx$$

Hence, substituting into (*) we get

$$\begin{aligned} \Gamma_{ijk}^{\theta_0} &= \int_X \partial_{ij}^2 \ln p(x, \theta) \cdot \partial_k \ln p(x, \theta) \cdot p(x, \theta_0) dx \\ &= E_{\theta_0} [\partial_{ij}^2 l \cdot \partial_k l] \end{aligned}$$

Q.E.D.

Examples.

The form of the Christoffel symbol is interesting. Suppose we consider our parameter space M as being embedded in the space $C^\infty(X, \mathbb{R})$. Where X is the sample space. Consider the function

$$E_{\theta_0}[\cdot, \cdot]: \Psi \rightarrow \mathbb{R}$$

where Ψ is the subset of $(C^\infty(X, \mathbb{R}))^2$ on which $E_{\theta_0}[\cdot, \cdot]$ is well defined. Now although $E_{\theta_0}[\cdot, \cdot]$ defined on this set Ψ is not necessarily positive definite we can still use it to project elements of the tangent space of Ψ onto the tangent space of M , i.e. the function

$$\begin{aligned} E_{\theta_0}[\cdot, \cdot]: T\Psi \times TM &\rightarrow \mathbb{R} \\ (\alpha, t) &\rightarrow E_{\theta_0}[\alpha, t] \end{aligned}$$

can be used to define the orthogonal directions to TM by the space

$$\{\alpha \mid E_{\theta_0}[\alpha, t] = 0 \quad \forall t \in TM\}.$$

If we have a curved submanifold, N , isometrically embedded in a larger flat space, M , then the connection on N is defined by the embedding in the following way. The connection is simply the way in which we define the derivatives of the tangent fields on N . Let $X(t)$ be such a vector field then its derivative with respect to t will not in general lie in the tangent space to N , but since M is flat it will lie in TM . Hence the connection on N is simply a way of projecting the tangent space of M down to that of N .

By analogy with the above definition of embedding connection we can view the Levi-Civita connection of $g_{ij}^{\theta_0}$ as an embedding connection of the parametric family in $C^\infty(X, \mathbb{R})$.

The definition of this embedding connection is simply to define the Christoffel symbols to be the projection under $E_{\theta_0}[\cdot, \cdot]$ of the first derivatives of the tangent vectors $\{\partial_i l\}$ back down into the tangent space itself. Thus

$$(\Gamma^{\text{embedding}})_{ijk} = E_{\theta_0}[\partial/\partial\theta_i(\partial_j l), \partial_k l]$$

So we see that $(\Gamma^{\text{embedding}})_{ijk} = \Gamma_{ijk}^{\theta_0}$.

Examples.

This gives us an interpretation of the g^{θ_0} metric as the one induced by the embedding of M into a larger function space.

Now note also that, if we are using this interpretation then the possibility arises that $p(x, \theta_0)$, the true distribution does not lie on the manifold M at all. As long as $p(x, \theta_0)$ is sufficiently close to M then g^{θ_0} will still be a metric. Hence this opens up more possible structures using the preferred point expected information. Again we shall not develop this idea here.

6.3 Example of g^{θ_0} in an exponential family.

We shall now work out explicitly what the g^{θ_0} metric is in an exponential family.

We define

$$p(x, \theta) = \exp\{t^i(x)\theta_i + a(x) - b(\theta)\} \quad (i = 1, \dots, p)$$

to be our exponential family M .

Thus,

$$l(\theta) = \ln p(x, \theta) = t^i(x)\theta_i + a(x) - b(\theta)$$

hence the tangent space to M is spanned by

$$\partial_i l(\theta) = t^i(x) - \partial_i b(\theta).$$

We can, therefore, calculate the coefficients of g^{θ_0} in the $\theta^1, \dots, \theta^p$ coordinates.

$$\begin{aligned} g^{\theta_0}_{ij} &= E_{\theta_0}[\partial_i l(\theta) \cdot \partial_j l(\theta)] \\ &= E_{\theta_0}[(t^i(x) - \partial_i b(\theta)) \cdot (t^j(x) - \partial_j b(\theta))] \end{aligned}$$

The following calculation is due to Dr. F. Critchley. Note that

Examples.

$$\int_{\mathbf{x}} \exp\{t^i(\mathbf{x})\theta_i + a(\mathbf{x}) - b(\theta)\} d\mathbf{x} = 1$$

By differentiating with respect to θ_i we get

$$\int_{\mathbf{x}} (t^i(\mathbf{x}) - \partial_i b(\theta)) \exp\{t^i(\mathbf{x})\theta_i + a(\mathbf{x}) - b(\theta)\} d\mathbf{x} = 0$$

Thus,

$$E_{\theta}[t^i(\mathbf{x})] = \partial_i b(\theta).$$

If we denote $E_{\theta}[t^i(\mathbf{x})]$ by $\mu^i(\theta)$ we see that,

$$\begin{aligned} \partial_i \ln p(\theta) &= t^i(\mathbf{x}) - \mu^i(\theta) \\ &= (t^i(\mathbf{x}) - \mu^i(\theta_0)) - (\mu^i(\theta) - \mu^i(\theta_0)) \end{aligned}$$

and so,

$$E_{\theta_0}[\partial_i \ln p(\theta) \partial_j \ln p(\theta)] =$$

$$\begin{aligned} &E_{\theta_0}[(t^i(\mathbf{x}) - \mu^i(\theta_0))(t^j(\mathbf{x}) - \mu^j(\theta_0))] - E_{\theta_0}[(t^i(\mathbf{x}) - \mu^i(\theta_0))(\mu^j(\theta) - \mu^j(\theta_0))] \\ &- E_{\theta_0}[(\mu^i(\theta) - \mu^i(\theta_0))(t^j(\mathbf{x}) - \mu^j(\theta_0))] + E_{\theta_0}[(\mu^i(\theta) - \mu^i(\theta_0))(\mu^j(\theta) - \mu^j(\theta_0))] \end{aligned}$$

Note that the metric g^{θ_0} has a particularly neat form:

$$\begin{aligned} g^{\theta_0} &= E_{\theta_0}[(t^i(\mathbf{x}) - \mu^i(\theta_0))(t^j(\mathbf{x}) - \mu^j(\theta_0))] + (\mu^i(\theta) - \mu^i(\theta_0))(\mu^j(\theta) - \mu^j(\theta_0)) \\ &= I_{\theta_0} + (\mu(\theta) - \mu(\theta_0))(\mu(\theta) - \mu(\theta_0))^T \end{aligned}$$

where I_{θ_0} is the Fisher information and T denotes transpose.

Examples.

6.4. Maximum Likelihood Estimate geometry.

In the previous section we generalised the Fisher information

$$E_{\theta}[\partial_i \ln p(x, \theta) \cdot \partial_j \ln p(x, \theta)]$$

to a preferred point version

$$g^{\theta_0}(\theta) = E_{\theta_0}[\partial_i \ln p(x, \theta) \cdot \partial_j \ln p(x, \theta)].$$

Often, the Fisher information is written as $E_{\theta}[-\partial^2_{ij} \ln p(x, \theta)]$. However, the natural preferred point generalisation of this, $E_{\theta_0}[-\partial^2_{ij} \ln p(x, \theta)]$, is not a metric. The reason for this is that it does not transform correctly under a change of coordinates. It is only a tensor at points where $\partial_i \ln p(x, \theta) = 0$, there the correct change of basis formula holds. In detail we see that if we change from θ -coordinates to ψ -coordinates then

$$\frac{\partial l}{\partial \psi_i} = \frac{\partial \theta_k}{\partial \psi_i} \frac{\partial l}{\partial \theta_k}$$

so,

$$\frac{\partial^2 l}{\partial \psi_i \partial \psi_j} = \frac{\partial \theta_p}{\partial \psi_i} \frac{\partial \theta_q}{\partial \psi_j} \frac{\partial^2 l}{\partial \theta_p \partial \theta_q} + \frac{\partial l}{\partial \theta_p} \frac{\partial^2 \theta_p}{\partial \psi_i \partial \psi_j} \quad (*)$$

Thus we can see from these equations that it is appropriate to use the hessian of a function as a metric as long as the condition $\partial_i \ln p(x, \theta) = 0$ holds. The observed geometry of Barndorff-Nielsen uses

$$\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}(\hat{\theta})$$

as a metric. This works because it is evaluated at the maximum likelihood estimate $\hat{\theta}$ where the condition on the first derivative will hold.

The observed geometry has the property that it is dependent on knowing some ancillary constant (see Chapter 1). If we wish to have an ancillary free

Examples.

geometry we must integrate out the dependence on the ancillary statistics. Let us define

$$A(\theta) = \{x \in X \mid \text{the maximum likelihood estimate of } x \text{ is } \theta\}$$

where X is the space of repeat samples for some i.i.d. distribution (see Chapter 1). Then we define

$$(g_{ml}^{\theta_0}(\theta))_{ij} = \int_{A(\theta)} -\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \cdot p(x, \theta_0) dx$$

6.4.1 Lemma: $g_{ml}^{\theta_0}$ is a preferred point metric in an open neighbourhood of θ_0 .

Proof. We first show that $g_{ml}^{\theta_0}$ transforms as a tensor. If ψ are the new coordinates. Then the change of basis will be given by;

$$\begin{aligned} (g_{ml}^{\theta_0}(\psi))_{ij} &= \int_{A(\theta)} -\left[\frac{\partial \theta_p}{\partial \psi_i} \frac{\partial \theta_q}{\partial \psi_j} \frac{\partial^2 l(\theta)}{\partial \theta_p \partial \theta_q} + \frac{\partial l}{\partial \theta_p} \frac{\partial^2 \theta_p}{\partial \psi_i \partial \psi_j} \right] \cdot p(x, \theta_0) dx \\ &= \frac{\partial \theta_p}{\partial \psi_i} \frac{\partial \theta_q}{\partial \psi_j} \int_{A(\theta)} -\frac{\partial^2 l(\theta)}{\partial \theta_p \partial \theta_q} p(x, \theta_0) dx \\ &= \frac{\partial \theta_p}{\partial \psi_i} \frac{\partial \theta_q}{\partial \psi_j} g_{ml}^{\theta_0}(\theta) \end{aligned}$$

since at the m.l.e. $\partial_i \ln p(x, \theta) = 0$ for all i . Therefore the formula obeys the correct change of basis rule for a metric.

We need to show that $g_{ml}^{\theta_0}$ is positive definite to complete the proof. This follows because θ is the maximum likelihood estimate. Hence the Hessian of the likelihood is -ve definite.

Q.E.D.

6.4.2 Note: We point out that there is a non-preferred point version of this metric given by

Examples.

$$\int_{A(\theta)} -\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \cdot p(x, \theta) dx$$

We shall not however follow this up here in any detail.

The maximum likelihood metric tells us the expected (from a θ_0 viewpoint) precision of the maximum likelihood estimator. Hence, in a sense, the geometry is a geometry which reflects the expected behaviour of the maximum likelihood estimate. So all geometric or statistical results from this geometry will reflect the interaction of the manifold with the estimate $\hat{\theta}$. From this viewpoint we can see that in comparing the $g_{ml}^{\theta_0}$ -geometry with, say, the previous g^{θ_0} -geometry or even the Fisher geometry we are assessing the performance of the estimator on our particular manifold. These ideas are very informally expressed here. We have not been able to produce any direct results from them yet. However we consider them in Chapter 8 where we have listed a series of intuitive ideas and conjectures on the relationship between this geometry and the behaviour of the maximum likelihood estimator.

6.5 Application to an Exponential Family.

As in section 6.3 we shall work out explicitly $g_{ml}^{\theta_0}$ on the exponential family,

$$p(x, \theta) = \exp\{t^i(x)\theta_i + a(x) - b(\theta)\}.$$

Then if $\hat{\theta}$ is the maximum likelihood estimate for the sample x , we have

$$\partial_i \ln p(x, \hat{\theta}) = 0.$$

Thus
$$t^i(x) - \partial_i b(\hat{\theta}) = 0,$$

or
$$t^i(x) = \partial_i b(\hat{\theta}).$$

Further we see that,

Examples.

$$\partial_{ij}^2 \ln p(x, \theta) = \partial_{ij}^2 b(\hat{\theta})$$

Therefore we see that the metric in θ -coordinates is given by

$$\begin{aligned} (g_{ml}^{\theta_0}(\theta))_{ij} &= \int_{A(\theta)} -\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \cdot p(x, \theta_0) dx \\ &= \int_{A(\theta)} -\partial_{ij}^2 b(\theta) \cdot p(x, \theta_0) dx = -\partial_{ij}^2 b(\theta) \cdot \int_{A(\theta)} p(x, \theta_0) dx \\ &= -m_{\theta_0} A(\theta) \partial_{ij}^2 b(\theta) \end{aligned}$$

Where $m_{\theta_0} A(\theta)$ is the measure of the ancillary manifold in X which maps onto θ under the maximum likelihood estimate. This is measured of course with the preferred point θ_0 -measure.

6.5.1 Lemma: $g_{ml}^{\theta_0}$ is conformally equivalent to the Fisher metric in the full exponential family case.

Proof. By definition the Fisher metric $g(\theta)$ is given by

$$\begin{aligned} g_{ij}(\theta) &= E_{\theta}[\partial_i \ln p(x, \theta) \cdot \partial_j \ln p(x, \theta)] \\ &= E_{\theta}[-\partial_{ij}^2 b(\theta)] \\ &= -\partial_{ij}^2 b(\theta). \end{aligned}$$

Thus we have that;

$$g_{ml}^{\theta_0}(\theta) = m_{\theta_0} A(\theta) \cdot g(\theta)$$

Therefore the two metrics are conformally equivalent.

Q.E.D.

Lemma 6.5.1 has a couple of important consequences. The first is that it gives a nice interpretation of the $g_{ml}^{\theta_0}$ metric as being a scaling of the Fisher metric,

Examples.

where the scaling depends on the likelihood that if θ_0 is the true parameter we actually use θ as our maximum likelihood estimate.

The second consequence is contained in the following lemma.

6.5.2 Lemma: $g_{ml}^{\theta_0}$ and the Fisher metric are first order asymptotically equivalent in the sense of chapter 6

Proof: This follows directly from lemma 6.5.1 and theorem 4.3.5.

6.6. The Expected Log Likelihood Surface.

We now come to an excellent explicit example of a preferred point manifold. Although it is an example from statistics it does give a good geometric feel for the properties of a preferred point geometry.

We define the θ_0 -expected log likelihood manifold (E.L.L.(θ_0)) to be the graph defined by

$$\text{E.L.L.}(\theta_0) = (\theta_1, \dots, \theta_p, E_{\theta_0}[\ln p(x, \theta)]) \quad (*)$$

We look at the preferred point geometry as being induced by the embedding of E.L.L.(θ_0) in \mathbb{R}^{p+1}

This preferred point geometry is a little different from the others we have been studying up to now. There is, for example, much less relevance to inference or information theory and also it is not a direct generalisation of Amari's structures.

6.6.1. Lemma: If (*) defines an embedding of E.L.L.(θ_0) in \mathbb{R}^{p+1} . Then the induced metric on E.L.L.(θ_0) is defined in the θ -coordinate system by

$$(g^{\theta_0}_{ell})_{ij} = \delta_{ij} + E_{\theta_0}[\partial_i \ln p(x, \theta)] E_{\theta_0}[\partial_j \ln p(x, \theta)]$$

where $\delta_{ij} = 1$ if $i=j$
 $\quad \quad = 0$ otherwise.

Examples.

Proof. At a point $(\theta_1, \dots, \theta_p, E_{\theta_0}[\ln p(x, \theta)])$ on the E.L.L. manifold the basis for the tangent space of E.L.L. which is induced by the θ -coordinates is

$$\{(1, 0, \dots, 0, \partial_1 E_{\theta_0}[\ln p(x, \theta)]), (0, 1, 0, \dots, 0, \partial_2 E_{\theta_0}[\ln p(x, \theta)]), \dots, (0, 0, \dots, \partial_p E_{\theta_0}[\ln p(x, \theta)])\}$$

We shall call this basis $\{t_1, \dots, t_p\}$.

By the regularity conditions of Chapter 1 section 3 we see that

$$\partial_i E_{\theta_0}[\ln p(x, \theta)] = E_{\theta_0}[\partial_i \ln p(x, \theta)].$$

We are going to use the metric on E.L.L. which is induced by the embedding in standard Euclidean space. Thus the i - j th term of the induced metric is given by the Euclidean inner product of the i th and the j th element of the above basis [do Carmo]. If we denote the standard Euclidean inner product by $\langle \cdot, \cdot \rangle$ we see that

$$\begin{aligned} (g^{\theta_0}_{ell})_{ij} &= \langle t_i, t_j \rangle \\ &= \delta_{ij} + \partial_i E_{\theta_0}[\ln p(x, \theta)] \cdot \partial_j E_{\theta_0}[\ln p(x, \theta)] \\ &= \delta_{ij} + E_{\theta_0}[\partial_i \ln p(x, \theta)] E_{\theta_0}[\partial_j \ln p(x, \theta)] \end{aligned}$$

Q.E.D.

We can now see how this preferred point metric differs from the others we have been studying. The metric at θ_0 is not given by I_{θ_0} the Fisher information at θ_0 . Instead it is just the identity matrix. Thus the metric, at least from this point of view would seem to have little to do with any inference based metric.

We shall now look at some of the properties of E.L.L. (θ_0)

6.6.2. Lemma: E.L.L. (θ_0) has a maximum at θ_0 .

Proof. We note that for all $i \in \{1, 2, \dots, p\}$

$$\partial_i E_{\theta_0}[\ln p(x, \theta)]|_{\theta=\theta_0} = E_{\theta_0}[\partial_i \ln p(x, \theta)]|_{\theta=\theta_0}$$

Examples.

$$\begin{aligned}
 &= \int_{\mathbf{x}} \frac{\partial_i p(\mathbf{x}, \theta_0)}{p(\mathbf{x}, \theta_0)} \cdot p(\mathbf{x}, \theta_0) \cdot d\mathbf{x} \\
 &= \int_{\mathbf{x}} \partial_i p(\mathbf{x}, \theta_0) d\mathbf{x} \\
 &= 0
 \end{aligned}$$

This is a maximum since

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} E_{\theta_0} [\ln p(\mathbf{x}, \theta)]|_{\theta=\theta_0} = -\text{Fisher information at } \theta_0$$

which is positive definite by assumption.

Q.E.D.

The connection induced by $g^{\theta_0}_{\text{ell}}$ has the Christoffel symbol given by the following lemma.

6.6.3. Lemma: In the θ -coordinate system the connection induced by $g^{\theta_0}_{\text{ell}}$ has the Christoffel symbols given by

$$(\Gamma^{\theta_0}_{\text{E.L.L.}})_{ijk} = E_{\theta_0} [\partial^2_{ij} \ln p(\mathbf{x}, \theta)] E_{\theta_0} [\partial_k \ln p(\mathbf{x}, \theta)].$$

Proof. We simply apply the formula

$$\Gamma^{\theta_0}_{ijk} = \frac{1}{2} \left(\frac{\partial g^{\theta_0}_{kj}}{\partial \theta_i} + \frac{\partial g^{\theta_0}_{ki}}{\partial \theta_j} - \frac{\partial g^{\theta_0}_{ij}}{\partial \theta_k} \right)$$

Q.E.D.

Examples.

6.7 The Exponential Family Case.

We shall now look at this metric in the exponential family case as before. Letting

$$p(x, \theta) = \exp\{t^i(x)\theta_i + a(x) - b(\theta)\}.$$

6.7.1 Lemma: In the exponential family case

$$(g^{\theta_0}_{\text{ell}})_{ij} = \delta_{ij} + (\mu^i(\theta) - \mu^i(\theta_0))(\mu^j(\theta) - \mu^j(\theta_0))$$

where $\mu^i(\theta) = E_{\theta}[t^i(x)] = \partial_i(b(\theta))$.

Proof. From the basic definition

$$\ln p(x, \theta) = t^i(x)\theta_i + a(x) - b(\theta)$$

Therefore,

$$\begin{aligned} E_{\theta_0}[\ln p(x, \theta)] &= E_{\theta_0}[t^i(x)\theta_i + a(x) - b(\theta)] \\ &= \theta^i E_{\theta_0}[t^i(x)] + E_{\theta_0}[a(x)] - b(\theta) \end{aligned}$$

Thus,

$$\begin{aligned} (g^{\theta_0}_{\text{ell}})_{ij} &= \delta_{ij} + (E_{\theta_0}[t^i(x)] - \partial_i b(\theta_0))(E_{\theta_0}[t^j(x)] - \partial_j b(\theta_0)) \\ &= \delta_{ij} + (\mu^i(\theta) - \mu^i(\theta_0))(\mu^j(\theta) - \mu^j(\theta_0)) \end{aligned}$$

Q.E.D.

By comparing this formula to that for the g^{θ_0} -metric in section 6.3 we notice an interesting similarity. We see that we get

$$g^{\theta_0} = I_{\theta_0} + (\mu(\theta) - \mu(\theta_0))(\mu(\theta) - \mu(\theta_0))^T$$

Because of this likeness we can propose a new metric, particularly for the exponential family, which is related to $g^{\theta_0}_{\text{ell}}$. Instead of embedding E.L.L. in \mathbb{R}^{p+1} with the standard metric we can embed it in \mathbb{R}^{p+1} with the metric

Examples.

$$\left(\begin{array}{c|c} I_{\theta_0} & 0 \\ \hline 0 & 1 \end{array} \right)$$

This is still of course flat since it is constant metric. We have simply rescaled the first p -coordinates by the Fisher information at θ_0 . We shall call the induced metric on E.L.L., $\bar{g}_{\text{ell}}^{\theta_0}$.

6.7.2 Lemma: The exponential family with the metric $\bar{g}_{\text{ell}}^{\theta_0}$ is isometric to the exponential family with the metric g^{θ_0} via the identity map.

Proof. We simply calculate $\bar{g}_{\text{ell}}^{\theta_0}$ as in lemma 6.6.1. We find that

$$(\bar{g}_{\text{ell}}^{\theta_0})_{ij} = I_{\theta_0} + (\mu(\theta) - \mu(\theta_0)) \cdot (\mu(\theta) - \mu(\theta_0))^T$$

Q.E.D.

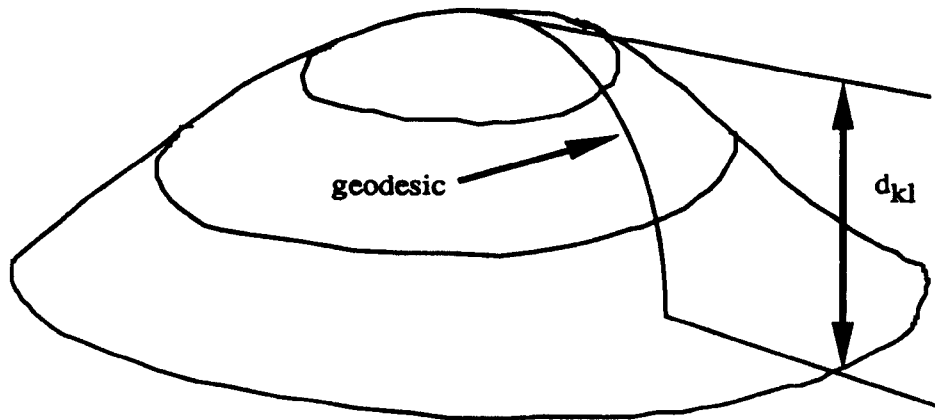
This isometry throws an interesting light on the relationship between the g^{θ_0} -metric and the divergence function given by

$$d_{\text{kl}}(\theta_0, \theta) = E_{\theta_0}[\ln p(x, \theta) - \ln p(x, \theta_0)]$$

We first recall that the g^{θ_0} -metric is the natural generalisation of the -1-connection of Amari. We also recall Amari's theorem concerning the relationship between the α -connections and the α -divergences. The -1-divergence is precisely d_{kl} , as above. Hence we would expect there to be a relationship between d_{kl} and g^{θ_0} . We can see what this is in explicit geometric terms if we consider g^{θ_0} as the embedding metric $\bar{g}_{\text{ell}}^{\theta_0}$ as we have shown we can above.

In the \mathbb{R}^{p+1} embedding we see that d_{kl} is simply the height between the two point θ_0 and θ . While the g^{θ_0} -geodesic distance is simply the $\bar{g}_{\text{ell}}^{\theta_0}$ -geodesic distance which is of course simply the shortest path from θ_0 to θ on our manifold.

Examples.



In the exponential case, therefore, we have found a direct geometric relationship between a divergence function and a corresponding geodesic distance. We can compare this result with both Amari's results on α -geodesics and α -divergences, and also our results in Chapter 5 on the relationship between divergence functions and preferred point geometries.

Chapter Seven

Preferred Point Geometric Theory.

7.1. Introduction.

In the two previous chapters we have seen some of the motivation for using a preferred point geometric structure as our basis for applying geometry to statistical theory. We have also studied various examples of preferred point geometries in statistics. The aim of this chapter is to develop some general theory of preferred point geometry and to show the similarities and differences with standard Riemannian geometry.

7.2. Comparison of Riemannian and Preferred Point Geometries.

Firstly there is an interesting relationship between the amount of structure of a preferred point geometry in comparison with that of a Riemannian structure. The complete preferred point structure $(M, g^{\theta_0}, \theta_0)$ has more geometric information than its Riemannian counterpart, (M, g) . This is because we can look upon $(M, g^{\theta_0}, \theta_0)$ as a family of Riemannian manifolds indexed by the manifold itself.

In applications, however we will often be looking at the preferred point geometry with θ_0 fixed at some value. Hence, in this way, we can say that the preferred point geometry is, under these common circumstances, precisely a Riemannian manifold and has exactly the same structure.

There is, however, another way of looking at the comparison. Suppose we have fixed θ_0 at some value. Then for certain purposes we are only going to be concerned with using geodesics which are measured from θ_0 rather than all the geodesics in the manifold. If we view the preferred point manifold in this way we see that it actually has a lot less structure than its Riemannian counterpart. We shall show this in detail in the next section where we consider the role played by curvature in a preferred point system.

7.3. Curvature in Preferred Point geometries.

The different aspects of a preferred point structure come out clearly in the consideration of the significance of curvature in a preferred point geometry context.

We will first review the role of curvature in standard Riemannian geometry. It has, of course, many uses but we shall recall just two of the most important ones.

First, as we saw in Chapter 6 it is an indicator of whether two Riemannian manifolds are isometric. For more details of this see Theorem 4.2.1. Essentially, if the manifolds have corresponding curvatures then they are isometric. So the first use of the curvature is that it distinguishes between two Riemannian manifolds.

If, in particular, the tensor is everywhere zero then we know that the manifold is isometric to flat space. Thus we can say that there exists an affine set of coordinates. Hence this is our second use of curvature and it is this property of the curvature which has been used in statistics by Amari and by Kass (see [Kass]).

Let us consider the first of these two uses in a preferred point context. The obvious and natural generalisation of curvature is that of *preferred point curvature*. This is defined as the curvature of the manifold (M, g^{θ_0}) once we have fixed on our choice of preferred point θ_0 . In Chapter 4 it was shown that in contrast with the Riemannian case two Statistical manifolds can have identical α -curvatures while not being isomorphic. There is an analogous result concerning the relationship between preferred point curvature and a preferred point manifold. We shall show the result later in the chapter but essentially it states the existence of two distinct preferred point manifolds which have the same θ_0 -curvature for each value of θ_0 .

This is an example of where the greater amount of structure in a preferred point geometry when we have not fixed the preferred point induces more complicated behaviour than that of the Riemannian case.

We shall now look at the second use of curvature in the preferred point context and we shall see how preferred point structures can sometimes exhibit simpler behaviour than Riemannian ones.

If (M, g) has non-zero Riemann-Christoffel curvature tensor then we know that there does not exist an affine set of coordinates. However if we are only concerned with geodesics from the fixed preferred point θ_0 then we see

Preferred Point Geometry Theory.

that we can always find an 'affine' type of coordinate system. By this we mean all the geodesics which pass through the preferred point are defined by linear maps in this coordinate system. This follows from the following well known theorem.

7.3.1 Theorem: (Existence of geodesic normal coordinate system)

In a Riemannian manifold (M, g^{θ_0}) there exists coordinate system (θ) such that the curves $\{ \theta_0 - t(\theta_1, \dots, \theta_p) \}$ are all geodesic for any vector $(\theta_1, \dots, \theta_p)$.

Proof. See [Sternberg]

Q.E.D.

The theorem states that there will always be a coordinate system such that the linear paths from the preferred point θ_0 are precisely the geodesics based at this point. Frequently it is only these geodesics based at the true parameter that we are interested in. Thus we can see from this use of the curvature that once we have fixed θ_0 we have in fact more freedom than the Riemannian model since we can find coordinates in which the geodesics which we are interested in are easy to calculate, no matter what the curvature is. This fact can greatly aid calculations since many general formulas reduce to their simplest form under these circumstances.

Although the curvature does not effect the existence of affine coordinates we can find an interpretation for it in terms of the measure of sets inside geodesic spheres based at θ_0 . We have already seen an example of this in the 'divergence' section of chapter 5. Loosely, the following theorem states that once you have fixed θ_0 , then the curvature tensor fixes the measure inside the geodesic spheres around θ_0 .

7.3.2 Theorem: Let the manifold M have two preferred point metrics g^{θ_0} and \tilde{g}^{θ_0} . Suppose also that around θ_0 the geodesic spheres of the two, centred on θ_0 agree. Further let us assume that at θ_0 the metrics agree. If we define

$$\phi: M \rightarrow M$$

by

$$\phi = (\exp_{g^{\theta_0}})^{-1}(\exp_{\tilde{g}^{\theta_0}})$$

Preferred Point Geometry Theory.

then ϕ is well defined in some neighbourhood of θ_0 . Further if in this neighbourhood, ϕ preserves the Riemann-Christoffel curvature tensor for the g^{θ_0} and \tilde{g}^{θ_0} metrics, then the measure of the g^{θ_0} -geodesic sphere of radius r is equal to that in the corresponding \tilde{g}^{θ_0} -geodesic sphere.

Proof. From Theorem 7.3.1 we see that there exists a neighbourhood in which the map ϕ is well defined.

Further by theorem 3.2.2 we also see that ϕ is an isometry. Hence the induced measures agree.

Q.E.D.

If we combine theorems 7.3.1. and 7.3.2. we see how curvature can play a different role in a preferred point structure to the one it plays in a Riemannian one. It can be seen as a function which describes the interrelationship between distance and the measures of sets around θ_0 rather than being an indicator of the existence of nice sets of coordinates.

This use of the curvature shows how using a preferred point analysis can sometimes actually produce simpler results than that of a Riemannian model.

7.4 Uncertainty.

In the previous section we have seen how, once the preferred point θ_0 has been fixed, a simplification of the geometric structure occurs. In this section we shall see that if the preferred point θ_0 is not fixed then we do have to understand the full structure of the preferred point geometry. We have already given an example of the greater richness of this structure in comparison to a Riemannian geometry. In many ways the strength of preferred point geometry in statistics is that the difference of information between the case when you know the true parameter and the case when you don't can be made explicit. Using the preferred point idea we can start to study the interrelationship between these two distinct cases. This section is just the start to such a study.

Let us start by considering how to use a preferred point system to study statistics. We would begin by picking θ_0 and then using the corresponding fixed geometry to study inference or efficiency in the ways made clear by Amari et al. However often we are trying to solve these problems when we do

Preferred Point Geometry Theory.

not know the true distribution exactly, or indeed at all. In fact we may often be faced with the problem of finding what the true parameter is. In this case we must, if we are to make use of preferred point geometry, make a choice of which point to fix on. Of course the geometric information which we are going to extract from our model will depend critically on θ_0 . We shall therefore define the differences in information generated by different choices of the preferred point as the *uncertainty of the preferred point geometry*.

One of the main examples where the uncertainty of a system is important occurs in the studying of an estimation problem. Very often since θ_0 , the true distribution, is not known it is assumed that $\hat{\theta}$, the maximum likelihood estimator is a good estimate of the true parameter. The standard derivation of the Wald Test for example uses the assumption [see appendix].

In our terms this assumption would mean working in the $(M, g^{\hat{\theta}})$ geometry rather than the (M, g^{θ_0}) geometry. We must therefore have ways of studying the difference in these two systems. It is, therefore, the uncertainty of the replacement of θ_0 with $\hat{\theta}$ that we must look at. We know that as the sample increases, then $\hat{\theta}$ converges in probability to θ_0 . We will therefore have convergence in the geometries $(M, g^{\hat{\theta}})$ and (M, g^{θ_0}) . It is one of the aims of the rest of this chapter to look at some of the ways in which these geometries differ and the statistical implications of the various forms of these differences.

We shall start by studying the ways in which we can understand the relationship between (M, g^{θ_1}) and (M, g^{θ_2}) . Clearly they are the same geometry if and only if the identity map

$$i:(M, g^{\theta_1}) \rightarrow (M, g^{\theta_2})$$

is an isometry. We can therefore make the following definition.

7.4.1. Definition: The preferred point geometry given by $(M, g^{\theta_0}, \theta_0)$ has *zero uncertainty* if for all $\theta_1, \theta_2 \in M$ the identity map

$$i:(M, g^{\theta_1}) \rightarrow (M, g^{\theta_2})$$

is an isometry.

Preferred Point Geometry Theory.

Thus we can trivially view Amari's expected geometry as a preferred point system with zero uncertainty. We shall now review a few fundamental concepts from Riemannian geometry which will be useful in the following work.

7.4.2 Definition and Notation: Let (M_1, g_1) and (M_2, g_2) be Riemannian manifolds, ∇_1 and ∇_2 be the associated Levi-Civita connections, and let

$$\varphi: (M_1, g_1) \rightarrow (M_2, g_2)$$

be a smooth map which has full rank. Let us also assume that

$$\text{Dim}(M_1) = \text{Dim}(M_2).$$

Then,

- (i) $\varphi^*(g_2)$ is the pullback of g_2 to M_1 .
- (ii) $\varphi^*(\nabla_2)$ is the pullback of the Levi-Civita connection of g_2 to (M_1, g_1)
- (iii) $\nabla_1 - \varphi^*(\nabla_2)$ is the second fundamental form (II.f.f) of the embedding.

For references to these definitions see [Spivak].

7.4.3. Definition: Using the same notation as above, except that now we assume that $\text{Dim}(M_1) \leq \text{Dim}(M_2)$. If the geodesics of (M_1, g_1) are mapped to the g_2 -geodesics in M_2 , then $\varphi(M_1)$ is said to be a *totally geodesic submanifold* of M_2 .

7.4.4 Lemma: If M_1 is the same manifold as M_2 and φ is the identity map, then the following are equivalent:

- (a) φ has zero second fundamental form.
- (b) $\varphi(M_1)$ is totally geodesic in (M_2, g_2) .
- (c) φ is an isometry.

Proof. Ref [Spivak].

Thus we have various criterion for the preferred point geometry to have zero uncertainty. Each of which can be useful in different contexts. However in most preferred point geometries we will not be in the zero uncertainty case. Thus we need to have ways of understanding the uncertainty and its statistical significance.

Preferred Point Geometry Theory.

We shall begin this study by looking at some maps between manifolds which preserve some but not all of the Riemannian structure. We have already studied the strongest condition on maps between Riemannian manifolds, i.e. isometries, and seen how they correspond to the zero uncertainty case. Therefore by studying maps which preserve less structure we shall see how they carry information about the uncertainty of a more general system.

7.4.5. Definition: With the notation of 7.4.2 we have the following definitions.

(C) $\varphi:(M, g^{\hat{\theta}}) \rightarrow (M, g^{\theta_0})$ is a *conformal map* if

$$g_1 = f(\theta)\varphi^*(g_2)$$

where f is a real valued smooth map from M_1 to \mathbb{R} .

(M) φ is an *measure preserving* map if the induced measure of $\varphi^*(g_2)$ equals the induced measure of g_1 .

Note: Condition (C) means that φ^* preserves the angle between tangent vectors, but not necessarily their lengths.

We recall that we have already seen both these conditions used earlier in this work. In Chapter 4 we showed that condition (C) is important to the study of the first order equivalences of exponential families (see Theorem 4.3.5). In Chapter 5 and 6 we have discussed the significance of the measure of a statistical family.

The first observation to make about (C) and (M) is that they can be looked at as complementary conditions in the following sense.

7.4.6 Lemma: Let (M_1, g_1) and (M_2, g_2) be equidimensional Riemannian manifolds, and let

$$\varphi:(M_1, g_1) \rightarrow (M_2, g_2)$$

be a smooth map of full rank. Then,

φ is an isometry if and only if (C) and (M) both hold.

Proof.

\Rightarrow :This is clear.

Preferred Point Geometry Theory.

\Leftarrow :If (C) holds then $g_1 = f(\theta)\varphi^*(g_2)$. Thus the induced measures are related by the following;

$$\det(g_1) = (f(\theta))^p \cdot \det(\varphi^*(g_2)).$$

Since (M) holds,

$$\det(g_1) = \det(\varphi^*(g_2)) \quad \text{almost everywhere.}$$

Hence $(f(\theta))^p = \pm 1$ almost everywhere.

Thus, by continuity and by the positivity of any metric, we see that $f(\theta) = 1$ for all θ .

In other words $g_1 = \varphi^*(g_2)$, i.e. φ is an isometry.

Thus (M) and (C) are complementary aspects of the idea of zero uncertainty in our preferred point manifolds. We can also use maps of type (M) or (C) to begin to classify various types of (non-zero) uncertainty in our theory.

7.4.8. Definition: We shall say that $(M, g^{\theta_0}, \theta_0)$ has *zero first order asymptotic uncertainty* if for all $\theta_0 \in M$ the metric g^{θ_0} is conformally equivalent to the Fisher metric.

7.4.9 Lemma: If M is an exponential family and $(M, g^{\theta_0}, \theta_0)$ has zero first order asymptotic uncertainty then, for any pair of preferred points, θ_0 and θ'_0 , the relevant geometries, (M, g^{θ_0}) and $(M, g^{\theta'_0})$, are first order equivalent (in the senses of definition 4.3.4.)

Proof. If the preferred point structure has zero first order asymptotic uncertainty then all the metrics are conformally equivalent to the Fisher metric. Hence they are all conformally equivalent to each other. Thus by Theorem 4.3.5 we have the result.

Q.E.D.

Therefore in an exponential family, if the preferred point structure has zero first order uncertainty, then up to the first order it does not matter which point you choose as the preferred point. This justifies replacing θ_0 the

Preferred Point Geometry Theory.

(unknown) true parameter with $\hat{\theta}$, the (known) maximum likelihood estimate. We could then make the calculations in the $\hat{\theta}$ -geometry and know that they would be correct to first order. Alternatively any calculation which only involved angles in the tangent space would be an exact one. An example of this would be the orthogonality calculations for which Amari shows first order efficiency.

7.4.10. Example: The expected maximum likelihood geometry $g^{\theta_0}_{ml}$ has zero first order asymptotic uncertainty on an exponential family. For a definition and proof of this see section 6.5.

How can we compare two geometries which both have zero first order asymptotic uncertainty? We can see that Lemma 7.4.6 gives us one measure of the difference. A preferred point system can have zero first order uncertainty and yet not have zero total uncertainty. We can use the difference between the induced volume measures for each geometry as a measure of the difference between the metrics. Thus we can propose a measure of the total uncertainty of a system with zero first order uncertainty to be

$$\text{Uncert}_{\Pi}(M, g^{\theta_0}, \theta_0) = (\det g^{\theta_0}) / (\det g),$$

where g is the Fisher metric.

7.4.11. Lemma: If $(M, g^{\theta_0}, \theta_0)$ is a system with zero first order uncertainty and $\text{Uncert}_{\Pi}(M, g^{\theta_0}, \theta_0) \equiv 1$ then the system has zero uncertainty.

Proof. By Theorem 7.4.6 we see that $(M, g^{\theta_0}, \theta_0)$ is the same as the standard Fisher metric structure. Hence it has zero uncertainty.

7.5 Local Measures of Uncertainty.

We have, so far, looked at cases where we have either zero uncertainty over the whole preferred point structure or where we have zero uncertainty up to first order. We now consider the problem of how to deal with the general case where there is non-zero uncertainty of some kind. The approach we follow is to look at various geometric objects and see how they

Preferred Point Geometry Theory.

behave as we alter the value of the preferred point. In this way, we produce various measures which can give indications of the reliability of the values attached to each of these geometric quantities. The statistical relevance of each of these is that they give some indication of how much the uncertainty of our knowledge of the true distribution matters in our analysis.

We shall now list the geometric quantities which we have previously considered. We shall then take each in turn and see how it behaves with a change in preferred point.

- (I) Measures of lengths of tangent vectors.
- (II) Measures of angles between paths or between tangent vectors. (This will of course include measures of orthogonality.)
- (III) The induced p -dimensional volume measure of a metric.
- (IV) Measures of the covariant curvature of curves in our manifold.

We have seen in previous chapters how (II) and (IV) have already been used in statistical theory, and we have given some indications about (I) and (III) although we have not explored these last two in the same detail as the others. Note however that all χ^2 distributions are measurements of type (I).

We now consider how the uncertainty of a preferred point geometry behaves at a single point. We will assume that we have some fixed point θ on the manifold. We shall however not assume that it is the preferred point, or true distribution. Rather we shall see how the geometry around θ changes as we select different preferred points. We shall often take as our measure of uncertainty of a particular geometric quantity its rate of change with respect to a change in preferred point. Clearly there are other measures possible and this section should be viewed as only as the start of a complete study of uncertainty.

One particular case will often be of interest. This is when our fixed point θ is the maximum likelihood estimate and we look at the uncertainty involved with making the assumption that the maximum likelihood estimate is the true parameter. The derivative of our geometric quantity will be evaluated at

Preferred Point Geometry Theory.

$\hat{\theta}$ itself. Since this is a common assumption in statistics it is important to understand the uncertainty under these circumstances.

7.5.1. Case (I): Lengths of tangent vectors.

Let $v \in TM_{\theta}$ be a tangent vector in the tangent space of some arbitrary, but fixed, point θ . We know that the length of v is a geometric quantity. If the preferred point metric is given by g^{θ_0} , then the length of v is given by

$$\|v\| = \sqrt{(g^{\theta_0}(v, v))}.$$

Let us write $g^{\theta_0}(\theta)$ as $g(\theta_0, \theta)$. We distinguish between the fixed (observed) point θ and the preferred point θ_0 by denoting θ_0 by ϕ . The preferred point metric will then be $g(\phi, \theta)$.

Thus we see that a measure of the uncertainty of the length is given by the rate of change of the squared length. For convenience we shall consider this quantity i.e., letting $\bar{\partial}_i = \partial/\partial\phi_i$, and $\partial_i = \partial/\partial\theta_i$

$$\begin{aligned}\bar{\partial}_i(\|v\|) &= \bar{\partial}_i(g(\phi, \theta)(v, v)) \\ &= \bar{\partial}_i(g_{jk}(\phi, \theta).v^j.v^k) \\ &= \bar{\partial}_i(g_{jk}.v^j.v^k) \\ &= \bar{\partial}_i g_{jk}.v^j.v^k \\ &= U_{ijk}v^j.v^k,\end{aligned}$$

where $U_{ijk} = \bar{\partial}_i g_{jk}$.

We must show that this definition is going to be a useful one. We first show how it behaves with respect to a change of coordinate system.

7.5.1.1 Lemma: If we define U_{ijk} as above then it obeys the following transformation rule as we change from θ to ψ coordinates.

Preferred Point Geometry Theory.

$$B_i^l(\phi).B_j^r(\theta).B_k^s(\theta).\bar{\partial}g_{rs}$$

where

$$B_r^s(\theta) = \frac{\partial \theta_s}{\partial \psi_r}(\theta)$$

Proof. Let the new coordinate system be $\psi(\theta)$. Then g^{θ_0} transforms as

$$B_j^r(\theta).B_k^s(\theta).g_{rs}$$

Thus by applying the chain rule we see that the full transformation rule is

$$B_i^l(\phi)\bar{\partial}(B_j^r(\theta).B_k^s(\theta).g_{rs}) = B_i^l(\phi).B_j^r(\theta).B_k^s(\theta).\bar{\partial}g_{rs}$$

Q.E.D.

7.5.1.2 Lemma: The object $U_{ijk}v^iv^j$ transforms like a tensor.

Proof. Since $g_{ij}(\phi,\theta)v^iv^j$ is independent of the coordinate system for each fixed ϕ we can view it simply as a function of ϕ . Thus its derivative $U_{ijk}v^iv^j$ behaves like a 1-tensor.

Q.E.D.

We can also look at the important special case where the observed point θ is also the preferred point, ϕ . This in fact is in many ways the most important case since if we are treating θ as the maximum likelihood estimate then this case means that the maximum likelihood estimate is treated as the true distribution. This is often the best that we can do.

7.5.1.3 Lemma: If $\theta = \phi$ then U transforms as a three tensor.

Preferred Point Geometry Theory.

Proof. We apply the transformation rule of 7.5.1.1 at $\theta = \phi$. This gives the result.

Q.E.D.

Suppose that, as in Chapter 5 the preferred point metric is given by a power series perturbation.

$$g^{\theta_0}_{ij} = g_{ij} + (\theta - \theta_0)^k T_{ijk} + (\theta - \theta_0)^k (\theta - \theta_0)^l S_{ijkl} + \dots \quad (*)$$

7.5.1.4 Lemma: If the preferred point metric is (*) then the uncertainty tensor U_{ijk} evaluated at θ_0 is given by the skewness tensor $-T_{ijk}$.

Proof. We see that by definition

$$\begin{aligned} U_{ijk} &= \bar{\partial}_i (g^{\phi}_{jk}) \\ &= \bar{\partial}_i (g_{jk} + (\theta - \phi)^l T_{jkl} + (\theta - \phi)^l (\theta - \phi)^m S_{jklm} + \dots) \\ &= T_{jki} + O((\theta - \phi)) . \end{aligned}$$

since g_{ij} , T , ... are all independent of ϕ .

The result follows if we evaluate U by setting $\theta = \phi$, and noting that T is symmetric.

Q.E.D.

This lemma therefore gives us an important insight into role of the skewness tensor. If we are making calculations at the point θ assuming that θ is the true parameter then the skewness tensor gives us a measure of the uncertainty involved in using that geometry. As we have stated before this is a very common situation in statistics, particularly if θ is the maximum likelihood estimate. Thus the size of the skewness tensor clearly gives us direct statistical information.

We can now see what U looks like in some of our examples of Chapter 6.

Preferred Point Geometry Theory.

7.5.1.5 Lemma: If we consider the metric given by $g^{\theta_0} = E_{\theta_0}[\partial_i l, \partial_j l]$, then for this case we have that

$$U_{ijk}(\phi, \theta) = E_{\theta_0}[\partial_i \ln p(x, \theta) \cdot \partial_j \ln p(x, \theta) \cdot \bar{\partial}_k \ln p(x, \phi)]$$

Proof. This is a simple calculation, using the regularity conditions of Chapter 1 and 2 to justify the swopping of partial derivative and integral signs.

Q.E.D.

7.5.1.6 Lemma: For the metric g^{θ_0} of section 2 in Chapter 6 we see that the tensor U is given by

$$E_{\theta}[\partial_i l \cdot \partial_j l \cdot \partial_k l],$$

i.e. Lauritzen's skewness tensor.

Proof. Again this is a simple calculation.

Q.E.D.

7.5.2 Case(II): Angles.

We shall now consider the case of the amount of uncertainty when measuring angles between tangent vectors. Again let θ be our arbitrary point on the manifold M . Also let ϕ denote the preferred point which we shall consider as varying while θ is fixed. We shall let v_1 and v_2 be any two tangent vectors in TM_{θ} . Now if we are working in the ϕ -geometry let us define

$$\mu_i = \frac{v_i}{\|v_i\|_{\phi}} \quad \text{for } i = 1, 2$$

Preferred Point Geometry Theory.

where $\|v\|_\phi = \sqrt{g^\phi(v,v)}$ is the norm on the tangent space at θ in the ϕ -geometry. So that for each i , μ_i is the normal tangent vector parallel to v_i . We use these to define the angle between the two vectors. If we let this angle be Θ then we know that by definition

$$\cos\Theta = g^\phi(\mu_1, \mu_2)$$

Thus we can consider as a measure of uncertainty of the angle between v_1 and v_2 , the measure

$$\begin{aligned} \bar{\partial}(\cos \Theta) &= \bar{\partial}(g^\phi(\mu_1, \mu_2)) \\ &= \bar{\partial}\left(\frac{g^\phi(v_1, v_2)}{\|v_1\|_\phi \cdot \|v_2\|_\phi}\right) \\ &= \bar{\partial}\left(\frac{g_{ij} \cdot v_1^i \cdot v_2^j}{\|v_1\|_\phi \cdot \|v_2\|_\phi}\right) \\ &= \frac{U_{ijk}(v_1, v_2) \cdot \|v_1\|_\phi^2 \|v_2\|_\phi^2 - g_{ij}(v_1, v_2)(U_{ijk}\|v_2\|_\phi^2 + U_{ijk}\|v_1\|_\phi^2)}{\|v_1\|_\phi^3 \cdot \|v_2\|_\phi^3} \end{aligned}$$

Thus we see that the uncertainty here is a function of U_{ijk} which was defined in the previous section (7.5.1). It is therefore in both cases (I) and (II) going to be a useful test of uncertainty.

We note that if the original measurement of the angle between v_1 and v_2 was a right angle then the formula for the uncertainty reduces to the simple

$$\frac{U_{ijk}(v_1, v_2)}{\|v_1\|_\phi \cdot \|v_2\|_\phi}.$$

Preferred Point Geometry Theory.

This is a result which applies directly to Amari's efficiency calculations of Chapter 2. In particular we recall Chapter 5 where his asymptotic results were reformulated in a preferred point geometry setting. Thus if we work under the assumption that the maximum likelihood estimate is the true parameter then again we find a direct interpretation of the Skewness tensor of Amari as the uncertainty that two paths cross orthogonally.

7.5.3 Case(III) Area.

We shall now consider how the total measure of a preferred point manifold changes under a change of preferred point. We have already seen how a preferred point metric induces p-dimensional measure on the manifold using the following formula:

$$\mu(\theta) = \det(g^{\phi}_{ij}) d\theta_1 \dots d\theta_p$$

We shall therefore consider the rate of change of $\det(g^{\phi}_{ij})$ as our measure of uncertainty, i.e.,

$$\bar{\partial} \det(g^{\phi}_{ij})$$

As in the previous chapters we must understand how this uncertainty statistic depends on the parametrisation. Since the determinant function is merely a polynomial function we see that $\det(g^{\phi}_{ij})$ is a well-behaved geometric object. In particular under the assumption that θ the observed point is the true parameter then it will be a tensor.

In general, however it is better to take the measure of uncertainty to be the related object

$$\bar{\partial} \mu(\theta) = \bar{\partial} (\det(g^{\phi}_{ij}) d\theta_1 \dots d\theta_p) \quad (\dagger)$$

The reason for looking at this is that $\mu(\theta)$ is independent of the parametrisation and thus is just a function on the manifold. Since $\bar{\partial}$ is a normal differential operator then $\bar{\partial} \mu$ will always be a tensor. As the 1-forms $d\theta_i$ are all independent of ϕ then equation (\dagger) reduces to

Preferred Point Geometry Theory.

$$\bar{\partial}\mu(\theta) = (\bar{\partial}\det(g^{\phi}_{ij})) d\theta_1 \dots d\theta_p.$$

7.5.3.1 Lemma: If the preferred point metric is given by $g^{\theta_0}_{ml}$ of section 6.4 and we are in the exponential family case

$$p(x, \theta) = \exp\{x^i \theta_i + a(x) - b(\theta)\}$$

then the uncertainty is of the form

$$\bar{\partial}m(\phi, \theta).m^{p-1}(\phi, \theta).\det \partial^2_{ij}b(\theta)$$

where,

$$m(\phi, \theta) = \int_{A(\theta)} p(x, \phi) dx$$

and $A(\theta)$ is the subset of the sample space which maps to θ under the maximum likelihood estimator.

Proof. In section 6.5 we see that in the exponential family case,

$$g^{\theta_0}_{ml}(\phi, \theta) = m(\phi, \theta).g(\theta)$$

where $g(\theta)$ is the Fisher information which is independent of the preferred point. Taking determinates we see that

$$\det g^{\theta_0}_{ml}(\phi, \theta) = m^p(\phi, \theta).\det g_{ij}$$

From Chapter 6 or by a simple calculation we see that

$$g_{ij} = \partial^2_{ij}b(\theta)$$

Thus we get the result from the independence of g with respect to ϕ .

Q.E.D.

Preferred Point Geometry Theory.

7.5.4 Case(IV): Geodesic curvature.

Amari shows how the geodesic curvature of curves in a manifold can be important statistically. Here we shall look at the uncertainty of measuring this using a preferred point geometry. The tool for measuring geodesic curvature is, of course, the Levi-Civita connection of the metric. Since this is dependent on the Christoffel symbols we shall look at their rate of change, with respect to ϕ , as a measure of the uncertainty involved. Thus we consider

$$\bar{\partial}\Gamma^{\phi}_{ijk}(\theta) \quad (\dagger\dagger)$$

as a choice for a measure of uncertainty of curvature.

Here, as with the Christoffel symbols themselves, we have to be careful about the behaviour under changes of coordinates. We recall that the Christoffel symbols are not tensors.

7.5.4.1 Lemma: The change of basis formula for $(\dagger\dagger)$ is given by the following

$$\bar{\partial}_t \tilde{\Gamma}^{\phi s}_{qr} = \bar{\partial}_v \Gamma^{\phi k}_{ij} \cdot B_t^v(\phi) \cdot B_q^i(\theta) \cdot B_r^j(\theta) \cdot \bar{B}_k^s(\theta)$$

where

$$B_i^j(\theta) = \frac{\partial \theta_j}{\partial \psi_i}(\theta) \quad \text{and} \quad \bar{B}_i^j(\theta) = \frac{\partial \psi_j}{\partial \theta_i}(\theta)$$

Proof. The change of basis formula for a Christoffel symbol is given by,

$$\tilde{\Gamma}^{\phi s}_{qr} = \Gamma^{\phi k}_{ij} \cdot B_q^i(\theta) \cdot B_r^j(\theta) \cdot \bar{B}_k^s(\theta) + \frac{\partial^2 \theta_i}{\partial \psi_q \partial \psi_r} \cdot \bar{B}_i^s(\theta)$$

hence the result comes from the chain rule and the fact that B is independent of the preferred point.

Q.E.D.

Preferred Point Geometry Theory.

We can see what $(\dagger\dagger)$ looks like in our examples of preferred point geometries. We shall take the case of the g^{θ_0} -geometry.

7.5.4.2 Lemma: In the g^{θ_0} -geometry of section 6.2 where

$$g^{\phi}(\theta)_{ij} = E_{\phi} [\partial_i \ln p(x, \theta) \cdot \partial_j \ln p(x, \theta)]$$

then the formula $(\dagger\dagger)$ reduces to

$$E_{\phi} [\partial^2_{ij} \ln p(x, \theta) \cdot \partial_k \ln p(x, \theta) \cdot \bar{\partial} \ln p(x, \phi)].$$

Proof. We have seen in section 6.3 that for this geometry the Christoffel symbols are:

$$\begin{aligned} \Gamma^{\phi}_{ijk} &= E_{\phi} [\partial^2_{ij} \ln p(x, \theta) \cdot \partial_k \ln p(x, \theta)] \\ &= \int_X \partial^2_{ij} \ln p(x, \theta) \cdot \partial_k \ln p(x, \theta) \cdot p(x, \phi) dX \end{aligned}$$

Assuming the usual regularity conditions (see chapter 1) that we can swop the derivative and integral signs, and noting the usual fact

$$\bar{\partial} p(x, \phi) = \bar{\partial} \ln p(x, \phi) \cdot p(x, \phi)$$

the result follows.

Q.E.D.

We again should point out the special case where we have $\theta = \phi$. In this case the uncertainty $(\dagger\dagger)$ is given by the fourth order cumulant

$$E_{\theta} [\partial^2_{ij} \ln p(x, \theta) \cdot \partial_k \ln p(x, \theta) \cdot \partial \ln p(x, \theta)].$$

Note that the uncertainty of the Christoffel symbol is a fourth order object rather than the third order objects we have seen when considering

Preferred Point Geometry Theory.

previous measures of uncertainty. This is a reflection of the higher order of the Christoffel symbols.

7.6. Uncertainty and Reparametrisation.

Let us consider a preferred point metric $g(\phi, \theta)$ as usual. If we consider the estimate, θ , as being fixed then we would have in general that g will vary with ϕ . The amount that it varies will, as we have seen above, depend on which coordinate system is being used. Thus we can ask the question, which coordinate system gives us the least variation over the whole choice of different possible preferred points? We can rephrase this question to see if by reparametrising can we reduce the dependence on the parameter ϕ ?

For fixed ϕ the function $g(\phi, \theta)$ is a metric on the manifold M . We shall take the reverse point of view, and study the function $g(\phi, \theta)$ with θ fixed and see how the two compare.

7.6.1 Notation: If we consider θ fixed we use the notation that

$$g(\phi, \theta) = g_\theta(\phi)$$

Now, although g_θ is a family of positive definite quadratic forms which are parameterised by M it, is not a metric and we must be careful about the subtle relationship between $g_\theta(\phi)$ and the preferred point metric $g^\phi(\theta)$. The difference of type between the two functions becomes clearer if we consider the spaces on which each one acts.

7.6.2 Definition: We consider the two functions given by $g_\theta(\phi)$ and $g^\phi(\theta)$. Then,

$$g^\phi(\theta): TM_\theta \times TM_\theta \rightarrow \mathbb{R} \quad \text{for all } \theta, \phi \in M$$

and

$$g_\theta(\phi): TM_\theta \times TM_\theta \rightarrow \mathbb{R} \quad \text{for all } \theta, \phi \in M$$

Preferred Point Geometry Theory.

where both the maps act in the obvious way.

We note that $g^\phi(\theta)$ acts on a different space for each value of θ , while $g_\theta(\phi)$ acts on a fixed space.

We consider their form under a change of basis. If we let the change of basis matrix at the point ψ be $B^i_j(\psi)$ then we have that

$$g^\phi(\theta)_{ij} \rightarrow B^l_i(\theta).B^k_j(\theta).g^\phi(\theta)_{lk}$$

and

$$g_\theta(\phi)_{ij} \rightarrow B^l_i(\theta).B^k_j(\theta).g_\theta(\phi)_{lk}$$

Note that if $g_\theta(\phi)_{ij}$ were a metric then it would have a change of basis formula like

$$g_\theta(\phi)_{ij} \rightarrow B^l_i(\phi).B^k_j(\phi).g_\theta(\phi)_{lk}$$

We see that the roles of θ and ϕ are not at all symmetric. There is however a form of duality between the two functions. It is not clear what the relationship is between this duality and the duality of Amari's α -connections.

We shall now see how reparametrisation effects g_θ and if we can choose coordinates to best effect.

From Definition 7.6.2 we see that, under reparametrisation of the manifold M , for each value of ϕ , the function $g_\theta(\phi)$ is acted on by the same change of basis formula, i.e.

$$B^l_i(\theta).B^k_j(\theta)$$

This is clearly independent of ϕ . Indeed, as θ is fixed, this change of basis is just a fixed quantity. Therefore we see that there is a large difference between the two cases of g_θ and g^ϕ . In the second case under the correct conditions (i.e. zero curvature) we can find a change of basis which will make the metric a constant. However in the dual case of g_θ we see that the only effect of a change of basis is to multiply g_θ by a fixed matrix. Thus if g_θ depends on ϕ in one basis then it will in all the others. The result of this is that a measure of the uncertainty of the preferred point system will only be trivially effected by a change of

Preferred Point Geometry Theory.

coordinates. Thus we can define a measure in any coordinate system and obtain an easy change of basis formula.

7.7 α -Flat Coordinate Systems in a Preferred Point Manifold.

In this chapter we have compared the structures of Riemannian and preferred point geometries. We have also looked at the theory of preferred point manifolds in their own right. In this section we will compare the structure of a Statistical manifold (following Lauritzen) to that of the related preferred point one.

In Chapter 5 it was shown that the non-metric connections of a Statistical manifold can be explained geometrically and generalised in terms of a larger preferred point geometry. For example it was shown how the -1 -curvature of Amari, at the true distribution, is simply the geometrically natural (Levi-Civita) curvature of a statistically natural preferred point manifold. We shall now look at the relationship between these two structures more closely.

One of the most interesting of Amari's results is that a full exponential family forms a -1 -flat manifold. Amari uses this result to greatly aid his calculations in many cases. However the statistical meaning of this result does not seem to have been fully explored. We shall see in this section what the condition of being -1 -flat implies about the preferred point manifold which generates the Statistical manifold.

Let $(M, g^{\theta_0}, \theta_0)$ be a preferred point manifold, we shall define the corresponding Statistical manifold by the following procedure. Recall that a Statistical manifold is made up essentially of a Riemannian manifold and a single non metric connection, the -1 -connection. We construct the Statistical manifold, which corresponds to our preferred point manifold, in two parts. The Riemannian manifold defined by $(M, g^{\theta}(\theta))$ is the Riemannian part of the statistical manifold. By this we mean at the point θ we take the metric which assumes that θ is the preferred point. The -1 -connection at θ is defined to be the Levi-Civita connection of the metric g^{θ} .

We recall from Chapter 5 that if we use this construction then Amari's Statistical manifold is precisely the one generated by the preferred point geometry g^{θ_0} studied in section 6.2.

Preferred Point Geometry Theory.

7.7.1 Theorem: The above procedure is well defined.

Proof. All we need to check is that $(M, g^\theta(\theta))$ is a Riemannian manifold, therefore it is sufficient to check that $g^\theta(\theta)$ is indeed a metric. Clearly it is a positive definite quadratic form on the correct tangent space. It is a smooth function of θ from the definition of the dependence of the metric on the preferred point. Hence all that we need to do is check that it transforms as a 2-tensor. This is clear since g^θ is a metric in the Riemannian manifold $(M, g^\theta(\psi))$, where ψ is any point in M .

Q.E.D.

Let $(M, g^{\theta_0}, \theta_0)$ be a preferred point geometry and let M^* be the corresponding Statistical manifold. We shall now see what the implications are for the preferred point manifold if M^* is -1-flat.

By the -1-flatness there exists a coordinate system on the manifold M which is -1-affine. We shall call this coordinate system ψ , say. Thus, on this coordinate system, the linear curve

$$\{ \lambda(\psi_1) + (1-\lambda)(\psi_2) \}$$

is precisely the -1-geodesic joining ψ_1 and ψ_2 . Now we shall see the geometric significance of this coordinate system in $(M, g^{\theta_0}, \theta_0)$. Pick a preferred point θ_0 , then the set of lines through θ_0 given by

$$\{ \lambda(\theta_0) + (1-\lambda)(\psi) \}$$

are not of course geodesics, but they all have zero curvature at the preferred point θ_0 .

Hence the line $\lambda(\psi_1) + (1-\lambda)(\psi_2)$ has the property that if you pick any point on it to be the preferred one then at that point the line will have zero curvature. We compare this to its property in the Statistical manifold where at each point it has zero -1-curvature.

We could therefore sum up this difference in roles by saying that in the preferred point case the line has zero curvature at each point whereas in the Statistical manifold case it has zero curvature everywhere. Certainly a vast difference.

7.8 Preferred Point Curvature and Classification.

In section 7.3 we stated that knowing the preferred point curvature at each point of a manifold was not enough to characterise the preferred point structure. This is an analogous result to that of Chapter 3 concerning Statistical manifolds. In this section we shall give an example of two distinct preferred point manifolds which have corresponding preferred point curvatures. This will prove the assertion in section 7.3.

We recall we defined the preferred point curvature of $(M, g^{\theta_0}, \theta_0)$ by the curvature of the manifold (M, g^{θ_0}) once we have fixed on our choice of preferred point θ_0 . Consider the following two examples of preferred point manifolds.

7.8.1 Examples. (i) $(\mathbb{R}^2 \setminus (0,0), \|(x,y)\| \cdot I, (x,y))$ where the manifold is the plane minus the origin and the metric is the identity matrix, I but scaled by the distance of the preferred point (x,y) from the origin, measured in the standard Euclidean norm.

(ii) $(\mathbb{R}^2 \setminus (0,0), I, (x,y))$, where we have the same manifold but the metric is independent of the preferred point (x,y) .

Example (ii) is clearly simply the standard Euclidean plane but here we are viewing it as preferred point manifold with zero uncertainty.

We see by inspection that for any fixed value of the preferred point (x,y) the curvature tensor in both cases will be identically zero. However these two example are certainly not isomorphic as preferred point metrics.

We shall return to this matter in Chapter 8.

7.9 Conclusions.

In this chapter we have just started to explore the theory of preferred point geometry. We have mostly concentrated on comparing the preferred point structure to both the Riemannian and the Statistical manifold structure of Lauritzen and Amari.

Preferred Point Geometry Theory.

There clearly remains a lot of work to do until this comparison is complete. There are two areas, at least, which we have not even touched on. The first is the significance, in the preferred point structure, of the duality on the non-metric connections of Lauritzen. The second, which relates to this, is the significance of the one parameter family of connections which Amari generates. Formally we can, of course, see that these aspects of Statistical geometry play a part in preferred point geometry theory. However it would be nice if we could produce a clear geometric interpretation of these constructions. We shall return to this discussion in the next chapter.

The main part of this chapter which relates to the theory of preferred point geometry in its own right are the sections on uncertainty. This seems to be an aspect of our structure which is new and has nothing corresponding to it in either Riemannian or Statistical manifold theory. The measures of uncertainty which we have produced are of a local nature, since they are based around the derivative of an object at a point. It is clear that different measures of uncertainty can be produced, either using higher derivatives or something which would give a more global measure. This again is work which remains to be done.

Chapter Eight Future Work.

8.1 Introduction.

In this chapter we shall detail some ways in which we think this work may be developed. Much of what will be described will not be completely worked through. Because of this we shall quite freely state conjectures which have not been proved and certainly all of the sections here are incomplete at best.

8.2. Some Questions Raised in this Work.

We start this chapter by looking at some questions which have been raised in earlier chapters. In particular we would like to look at some issues which we have dealt with either incompletely or not at all. We shall start with a review of what was achieved in Chapter 5.

In the beginning of that chapter we raised some questions about the structure of Statistical manifold theory and we questioned why this same structure seemed to be useful in various parts of statistics which were apparently different. We tried to find a more natural geometric structure which unified the various applications and would point the way to further generalisations. The result of this brought about the definition of a preferred point geometry. We showed in what way this new structure was a generalisation of Lauritzen's Statistical manifold structure, and then we completed the chapter by showing how in fact a preferred point structure was a natural one for the study of divergence measures and a useful one for the study of asymptotic theory. We can now see what we have to do before the questions which we asked are fully answered. There seem to be two main areas where work needs to be done. The first is that we must show why a preferred point geometric structure is natural for *all* the applications which statistical geometry has been put to. The second issue is to explain all of the properties of

Future Work

Lauritzen's statistical manifold structure from a preferred point geometry viewpoint.

Taking the first question first, we shall explain the areas which we feel need working on and review the work so far. In Chapter 5 we stated that the three areas which so far have used the Lauritzen/Amari structure were asymptotic analysis, divergence functions and reparametrisation. The area of divergence functions is the one which we feel is covered most thoroughly by our current work. We have shown that the preferred point set up is a natural framework for many divergence functions and have extended the paper of Rao connecting differential geometry and divergence functions. In the area of asymptotic analysis we feel we have been partly successful. We have shown that the results of Amari fit very naturally into a particular preferred point structure. What needs to be done is to find some more basic reasons why the results that we do have are correct. It is not clear to us that, if we started with a preferred point structure, an application to asymptotic analysis would have been a natural one without the work of Amari to start with. We need compelling geometric reasons which comes from the nature of the asymptotic analysis itself which will force the preferred point structure upon us. In section 8.5 we shall indicate a possible solution to this question although the work there is very incomplete. Also the section 8.3 will be relevant to this discussion. It would be interesting to see if the preferred point geometry actually extended Amari's results to higher order asymptotic theory than third order.

It is the last of these areas in which most work needs to be done. We have not tackled at all the way in which preferred point geometry is appropriate to the question of reparameterisation to which Amari applies the α -connections. We quoted his theorem on this subject in Theorem 2.9.4. He finds statistical meaning for affine coordinates for values of $\alpha = 0, \pm 1, \pm(1/3)$. Thus we have the question of finding a preferred point geometry which explains each of these results. We do have, as we have seen, preferred point geometries which correspond to these values of α . However, their application to the reparametrisation issue is not clear

The second issue to which more work must be done is the question of how much the pure preferred point structure explains the structure of the Statistical manifold. One particular area that is interesting is the question of the duality of Lauritzen's structure. For example we know (see chapter 2) that an exponential family is 1-flat and a mixture family is -1-flat and that these two families are dual. There is no real understanding of why this is true or what the statistical implication of these facts are. An open question is, therefore, can preferred point geometry throw any light on these facts? We have explained in Chapter 7 the significance to preferred point geometry of α -affine coordinates,

it is not yet clear how this helps. Also it is not clear if the duality structure of the Statistical manifold extends to the preferred point structure in any way.

8.3. The Role of Geodesics.

In this section we wish to consider what the possible significance of geodesics are in a Riemannian structure when it is applied to statistical theory. In particular we shall consider their possible role in estimation theory, i.e. by using them to construct geodesic distances between points and using these distances as divergence type measures, as in Chapter 5. We shall not be able to construct a metric which is able to do that in an unambiguous way. In fact most of this section will show constraints on possible statistical metrics. The main point could be summed up by proposing that the existence of a metric which has statistical importance does not imply that the geodesics that it produces will have a *direct* statistical interpretation. We do hope however to list the conditions which need to be fulfilled if we are to obtain geodesics with such direct interpretations.

Although the idea of using some kind of geodesic distance on the manifold which corresponds to a parametric model is an intuitively nice one, there are some hidden assumptions and implications which must be noted in the statistical context. One example of these implications which we have already seen is the symmetry of geodesic distances. This is a consequence of the fact that all points on the statistical manifold have been given equal importance. This is known as the homogeneity of the manifold. We have seen that this is not necessarily correct or plausible statistically. Such considerations led us to turn to the added structure of a preferred point geometry. However the symmetry of the geodesic distance is not the only implication to be considered. We shall therefore review the definitions and structure of a Riemannian manifold and see what statistical considerations force upon us.

8.3.1. Linearisation and Geodesics.

The key to the idea of Riemannian geometry is that of the linearisation of the curved manifold at each point. In a Riemannian structure we note that *all* the local geometric information is stored in the tangent spaces. That is to say, if you understand what is happening on each tangent space then you know everything about the local geometry. The only information you don't know is the global topological information which, as we said in Chapter 1, we shall not

Future Work

be dealing with statistically anyway (see section 8.7). This localisation is done via a metric, we then have enough information to measure path lengths, curvature etc, in the standard ways via integration and solving differential equations. Thus the geodesic distance is determined entirely by the information put on the tangent spaces and by integration theory.

We see that, therefore there are two points to be considered. The first is that we can encode *all* our statistical information in the linearised manifold at each point. The second is that our concept of distance is based on one of path length i.e. our information can be meaningfully integrated along smooth curves. It is not clear that statistical information has this property. We must therefore consider both these points in the statistical context to which we want to apply the geodesics.

8.3.2. Statistical Implications.

If we are considering the relevance of geodesic distance to inference theory, then the previous two points both present difficulties.

The first point raises the question of why should (or how can) information relevant to inference be contained in the tangent spaces at each and every point of the manifold? To put the question another way, why is the inference question one that can be localised at each point? To study this issue we shall look at a couple of the most common views of the tangent space to a manifold and see that unfortunately neither viewpoint gives us a clear path to solving the problem. They do however bring to the fore the points that need answering.

I) The first interpretation of the tangent space is that it is the space of all tangents to all smooth paths on the manifold which go through the relevant base point. The reason why this is not a useful point of view for our inference problem is simply that there are no relevant smooth paths which the problem naturally brings up. We have, in general, just the null-hypothesis θ_0 and an estimate $\hat{\theta}$. We may in fact have a sequence of estimates $\hat{\theta}_n$. However we do not get a smooth path of estimates which the definition would seem to require. There is a different type of measurement going on here. The metric on each tangent space will give some indication of the rate of change of distance along the path, however the only place where the concept naturally arises is when measuring the rate of convergence to the true parameter. It is the lack of a natural differentiability in the convergence that cause the difficulty. Thus we

Future Work

have the situation where the definition of the metric is clear at the true parameter but much less so at other points on the manifold.

II) The second interpretation of the tangent space is that of the best linear approximation to the manifold at each point. This is, clearly, only a good approximation for some small differences from the base point. If this is going to be a useful concept there are a couple of points to be considered:

Firstly, to be useful in inference problems, the linearisation must be a sensible thing to do in this context, i.e., the inference problem must be easier to be solved in a linear manifold.

Secondly we must take into account the fact that the approximation of the manifold by a linear tangent space is only good for small differences, i.e. in the inference problem we must find a reason for only being interested in small difference from each point. As in I) there is an intuitive case for using small differences around the true parameter since we get convergence in probability to this point but for points away from the true parameter the question seem less clear cut.

The second point from the previous section also needs to be looked at in the statistical context. This is the fact that the geodesic distance is one based on path integrals. The point here is that having a direct statistical interpretation to the value of the metric at each point of the manifold does not insure a statistical interpretation to a path integral using this metric. To insure this the local statistical information would have to be *additive*, in some sense, to allow us to integrate. An example of where this does not happen is the following case. We can construct a metric which is based on the locally most powerful test which distinguishes $\hat{\theta}$ from $\hat{\theta} + \delta\hat{\theta}$. However it would not be appropriate to integrate these results over a path since each calculation of the metric is based on a different assumption of the true parameter. Thus, although the linearisation is meaningful the result is not integrable.

In the previous discussion although we haven't proved anything and we have been rather negative we have brought out the need for two principles to be considered for a direct interpretation of a geodesic statistic. These are *linearisation* and *integrability*. A direct interpretation of the geodesic is not of course a thing which we must have. Geodesics can just be considered tools which can produce useful results for the statistician, as we hope the survey of results in Chapters 1 and 2 have shown. Nevertheless we would prefer to achieve some sort of interpretation and this is one direction in which future work can go. While this section can be looked at as a list of the reasons why we have not achieved our aim, in the next couple of sections we will outline ways in which we feel progress may be made. The ideas proposed here, we have to say, have not been worked through and much research needs to be

done. The ideas do however indicate the direction that we feel our future work should take.

8.4. The Maximum Likelihood Estimator Geometry.

In this section we shall propose a geometric model for the behaviour of the maximum likelihood estimator under certain circumstances. This model is based on one of the preferred point metrics which is proposed in Chapter 6. We will not compare this model, either in a theoretical or an empirical way, with the actual behaviour of the m.l.e. This is work which remains to be done. What we will do is explain heuristically why the preferred point geometry appears to give the possibility of a good working model. In particular what we actually do is show how this model does fulfil the conditions of section 8.2 that we need to obtain a justification for the direct understanding of geodesic distance as a statistical quantity. Again we must stress that what we propose here is merely a possible model and finding ways of evaluating its performance remains a major open question.

We are going to consider the maximum likelihood estimator (m.l.e.) and try to present a geometry in which it lies in a natural way. The idea is to describe a geometry which fits the way that the m.l.e. behaves locally as repeated samples are taken. We shall then produce a stochastic model of its behaviour which lies in the global geometry which is defined by the preferred point metric that we have produced.

Apart from the natural interest in doing this for its own sake we hope in following up this idea, to illustrate the discussion in the last section. In this discussion we recalled that a use of geodesics in a statistical way implies that we are using two principles. The first is that of *linearisation* of our information, i.e., that whatever information we are interested in can be encoded easily into a local approximation of our manifold, for example the tangent space. The second principle which we noted is that of the *integrability* of the localised data, i.e., once we have localised our information the linearisation process is such that it enables us to sum or integrate it over, say a path, to regain the global information. The following discussion should help to give some context to this subject. As we said in the previous section these issues seem to be far from resolved here or in any of the literature.

In the process that we are trying to model we shall assume that we have the following situation. We are taking i.i.d. samples from a sample space

Future Work

X. The maximum likelihood estimates for the total sample after n observations will lie in a finite dimensional manifold of distributions given by

$$M = \{ p(x, \varphi) \mid \varphi \in \mathbb{R}^p \}.$$

We shall denote the estimate by $\hat{\varphi}(x(n))$. We have here that $x(n)$ is the total observation after n repeated samples which are generated by the true distribution given by the coordinate φ_0 . So we have the sample $x(n)$ and we shall take another sample to get $x(n+1)$. We shall see how the position of $\hat{\varphi}$ changes as we do this.

Assuming the regularity conditions of 1.3.2 we see that we have, by the definition of m.l.e., two equations corresponding to the sample sizes n and $n+1$

$$(C(n)) \quad \partial_i(\ln p(x_n, \hat{\varphi}(x(n)))) = 0 \quad \forall i \in \{1, \dots, p\}$$

and

$$(C(n+1)) \quad \partial_i(\ln p(x_{n+1}, \hat{\varphi}(x(n+1)))) = 0 \quad \forall i \in \{1, \dots, p\}$$

We shall use the notation that

$$\partial_i l(\cdot, \cdot) = \partial_i \ln p(\cdot, \cdot)$$

and that

$$\hat{\varphi}_n = \hat{\varphi}(x(n)). \quad n \geq 1$$

We then want to see the difference between $\hat{\varphi}_n$ and $\hat{\varphi}_{n+1}$. We shall define this difference to be

$$\delta \hat{\varphi}_n = \hat{\varphi}_{n+1} - \hat{\varphi}_n$$

We see, therefore, that we can write $(C(n+1))$ as

$$\begin{aligned} 0 &= \partial_i l(x_{n+1}, \hat{\varphi}_n + \delta \hat{\varphi}_n) \\ &= \partial_i (\sum_j \ln p(x^j, \hat{\varphi}_n + \delta \hat{\varphi}_n)) \end{aligned}$$

Future Work

$$\begin{aligned}
&= \partial_i \sum_{j=1}^{n+1} \ln p(x^j, \hat{\phi}_n + \delta \hat{\phi}_n) \\
&= \partial_i \sum_{j=1}^n \ln p(x^j, \hat{\phi}_n + \delta \hat{\phi}_n) + \partial_i \ln p(x^{n+1}, \hat{\phi}_n + \delta \hat{\phi}_n) \\
&= \partial_i \sum_{j=1}^n (\ln p(x^j, \hat{\phi}_n) + \partial_k \ln p(x^j, \hat{\phi}_n) \delta \hat{\phi}_n^k + O((\delta \hat{\phi}_n^k)^2)) \\
&\quad + \partial_i \ln p(x^{n+1}, \hat{\phi}_n + \delta \hat{\phi}_n) \\
&= \partial_i \sum_{j=1}^n (\ln p(x^j, \hat{\phi}_n)) + \partial_i \sum_{j=1}^n (\partial_k \ln p(x^j, \hat{\phi}_n) \delta \hat{\phi}_n^k) + O((\delta \hat{\phi}_n^k)^2) \\
&\quad + \partial_i \ln p(x^{n+1}, \hat{\phi}_n + \delta \hat{\phi}_n) \\
&= \partial_{ik}^2 \sum_{j=1}^n (\ln p(x^j, \hat{\phi}_n)) \delta \hat{\phi}_n^k + O((\delta \hat{\phi}_n^k)^2) + \partial_i \ln p(x^{n+1}, \hat{\phi}_n + \delta \hat{\phi}_n)
\end{aligned}$$

Here we have used equation (C(n)) to eliminate some of the constant terms. We can therefore write the above in matrix form as

$$(\partial_{ik}^2 \ln p(x_n, \hat{\phi}_n)) (\delta \hat{\phi}_n)^k = - \partial_i \ln p(x_{n+1}, \hat{\phi}_{n+1}) + O((\delta \hat{\phi}_n)^2) \quad (\dagger)$$

Because of the previous observations we can propose a model for the behaviour of the m.l.e. for large sample sizes.

Let us consider the following heuristic argument. We want to see how the m.l.e. $\hat{\phi}_n$ changes with n . If we now assume that n is large it is clear that the change from $\hat{\phi}_n$ to $\hat{\phi}_{n+1}$ will, with high probability, be a small one. We can see this from the Central Limit Theorem.

This is the essential part of the localisation argument. We are saying that if we are at *any* point of the manifold $\hat{\phi}$ then the movement to the next point will be small because the sample size is large enough. Thus it is important to know what happens in the small region around each point of the manifold. This is of course our second way of looking at the tangent space in section 8.2. Hence it is possible to encode information about the behaviour of the m.l.e. in the linearisation around $\hat{\phi}$. That is in the tangent space at $\hat{\phi}$.

We can use equation (\dagger) to give us the local behaviour of the estimator. We shall assume that $\delta \hat{\phi}_n$ is sufficiently small that we need only consider its

Future Work

first order terms. Then (†) tells us the relationship between $\delta \hat{\phi}_n$ and the tangent vector in that direction i.e., $\partial_i l(x_{n+1}, \hat{\phi}_{n+1})$. Thus we see that if we put the metric given by

$$n \int_{A(\hat{\phi}_n)} (\partial_{ij}^2 \ln p(x, \hat{\phi}_n)) p(x, \phi_0) dX \quad (*)$$

we have scaled the tangent vectors at each point to equal the expected jump in the m.l.e.

Thus our model is given by the following. We consider the manifold M with the preferred point metric given by (*). Then we model the m.l.e. by a point which moves by a Markov process on this manifold with its distribution at each point given by the distribution of the tangent vectors at each point scaled by the metric (*). Their distribution is calculated from the true distribution which is given by $p(x, \phi_0)$. Thus the point moves in a space whose geometry is preferred point.

We must consider the question of integrability. At first sight there appears to be a problem here. Since the metric is scaled at each point of the stochastic process by the sample size n , it would appear that there is some consistency problem since as n changes we are effectively working on a different Riemannian manifold, i.e. $(M, (n+1)g)$ rather than (M, ng) . Thus as the metric depends on the sample size it is not immediately clear on which Riemannian (or preferred point) manifold we should take our geodesics. However in this case the way that the Riemannian manifold is changing depends in a very simple way on the sample size, and in some ways this change is geometrically trivial. We can see that intuitively all that is happening is that the whole manifold is being scaled down with time. This scaling is just a simple linear one. It takes geodesics to geodesics and preserves angles. Thus geodesic spheres are taken to geodesic spheres. The main thing that changes is path length.

We shall look at the case of a point which is moving on a Riemannian manifold which itself is being scaled down with time, i.e. we are on the manifold given by $(M, (1/t)g)$ where t is time and we assume that $t > 1$. This will be the continuous version of our stochastic model where the sample size has been replaced by $1/t$. If we understand how a path moves in the (easily) time dependent geometry of this system we conjecture this will help to understand our stochastic model and thus the real situation. In the time dependent case we have the following very useful lemma which indicates that the time dependence is geometrically trivial.

Future Work

8.4.1 Lemma: If we measure the path length of a path $\gamma(t)$ in the time dependent manifold $(M, (1/t)g)$ by the natural generalisation of the standard formula given by:

$$I(\gamma) = \int_0^1 \sqrt{\frac{1}{t} g_{ij} \dot{\gamma}^i(t) \dot{\gamma}^j(t)} dt$$

then the curves which minimise $I(\gamma)$ are the geodesics of (M, g) .

Proof. We shall define

$$\varphi(\gamma^i(t)) = \frac{1}{t} \bar{\varphi}(\gamma^i(t)) = \sqrt{\frac{1}{t} g_{ij} \dot{\gamma}^i(t) \dot{\gamma}^j(t)}$$

We follow the standard calculus of variations argument which defines the geodesic equations.

Let $\gamma(t) + \varepsilon \omega(t)$ be a variation of the path which fixes the end points, then define

$$I(\varepsilon) = \int_0^1 \varphi(\gamma(t) + \varepsilon \omega(t)) dt$$

to be the length of the perturbed path. Then the first variation of this length is

$$\begin{aligned} \delta I(\varepsilon) &= \frac{d}{d\varepsilon} I(\varepsilon) = \frac{d}{d\varepsilon} \int_0^1 \varphi(\gamma + \varepsilon \omega) dt \\ &= \frac{d}{d\varepsilon} \int_0^1 \frac{1}{t} \bar{\varphi}(\gamma + \varepsilon \omega) dt \\ &= \int_0^1 \frac{1}{t} \left[\frac{\partial \bar{\varphi}}{\partial \gamma^i} \omega^i + \frac{\partial \bar{\varphi}}{\partial \dot{\gamma}^i} \dot{\omega}^i \right] dt \\ &= \int_0^1 \frac{1}{t} \left[\frac{\partial \bar{\varphi}}{\partial \gamma^i} \omega^i \right] + \frac{d}{dt} \left[\frac{1}{t} \frac{\partial \bar{\varphi}}{\partial \dot{\gamma}^i} \right] \omega^i dt \end{aligned}$$

Future Work

Thus the path will be a 'geodesic' if

$$\frac{1}{t} \left[\frac{\partial \bar{\phi}}{\partial \dot{\gamma}^i} - \frac{d}{dt} \frac{\partial \bar{\phi}}{\partial \dot{\gamma}^i} \right] + \frac{1}{t^2} \left[\frac{\partial \bar{\phi}}{\partial \dot{\gamma}^i} \right] = 0$$

or,

$$\frac{1}{t} \left[g_{ij} \ddot{\gamma}^j + \frac{\partial g_{ij}}{\partial \dot{\gamma}^k} \dot{\gamma}^j \dot{\gamma}^k - \frac{1}{2} \frac{\partial g_{ik}}{\partial \dot{\gamma}^i} \dot{\gamma}^j \dot{\gamma}^k - g_{ij} \dot{\gamma}^j \frac{\frac{d^2 s}{dt^2}}{\frac{ds}{dt}} \right] + \frac{1}{t^2} \frac{g_{ij} \dot{\gamma}^j}{\sqrt{g_{ij} \dot{\gamma}^j \dot{\gamma}^i}} = 0$$

where s is the path length. Then making the usual substitution for the Christoffel symbols we find that the condition reduces to,

$$\frac{1}{t} \left[g_{ij} \ddot{\gamma}^j + \Gamma_{jki} \dot{\gamma}^j \dot{\gamma}^k - \frac{g_{ij} \dot{\gamma}^j}{\frac{ds}{dt}} \left[\frac{d^2 s}{dt^2} - \frac{1}{t} \right] \right] = 0$$

Thus if we parametrise by a parameter such that

$$\left[\frac{d^2 s}{dt^2} - \frac{1}{t} \right] = 0$$

we see that the equation reduces to the standard one for geodesics on the non time dependent manifold.

Q.E.D.

The above lemma means that we can use a non-time dependent geometry (M, g) to give us a model of what happens in the time dependent case. In particular, the geodesic spheres will be the same point sets whether we use the time dependent or the independent version.

In the context of our model of the m.l.e. the above discussion means that in our model we have the integrability condition that we need since the

Future Work

dependence of the metric on the sample size turns out to be geometrically trivial.

Assuming that this is an accurate model the geodesic spheres centred at the true parameter in the preferred point metric will be sensible contours to take for our inference measure. The reason for this is simply that we are viewing the estimator as doing a random walk on this manifold thus each point on each geodesic sphere has an equal chance of being the m.l.e. after n -samples since the walk is random.

We have found a direct statistical interpretation of the geodesic statistic in this case. The problem of localisation has been interpreted as an assumption that the sample size is large enough so that the movement in the m.l.e. is small and the tangent space approximation is a good one. The integrability condition is a problem but is at least partially resolvable by Lemma 8.4.1. The question which remains open is the verification of the model either in a theoretical way or using numerical methods.

8.5 An Asymptotic Model.

We shall propose another model here which is similar in spirit to the previous one but introduces a different preferred point geometry. We have seen in Chapter 6 that the metric given by g^{θ_0} (see section 6.2) has good properties for its geodesic from an asymptotic point of view. We saw that the metric fitted well into Amari's asymptotic analysis and the geodesics produced third order efficient estimators. We said in section 7.2 that much work remains to be done on this metric, and to the application of preferred point geometry to asymptotic analysis. One open question is, do these geodesics in fact produce estimators which are better than third order efficient? If they don't, can we find a preferred point system which does improve on their efficiency?

These questions again might be answered if we could find a direct interpretation for the geodesics or the geodesic distance. The discussion of Section 8.3 again applies, and we would like to propose another model with similar motivation to that of section 8.4 which might be able to give this interpretation.

We again have the two questions of integrability and linearisation to consider here. We shall in fact tackle both of them in the same way that we did in section 8.4. We shall work in a model which is doing i.i.d. sampling with sample size n . We shall not this time specify which estimator we are using but think of some idealised efficient estimator. Again we shall assume that the

Future Work

sample size is large enough so that the distance moved by the estimate for size n and that for $n+1$ is so small that we can understand it by dealing with the local linearisation i.e. the tangent space. We then look at the distribution of the vectors in the tangent space and we work under the assumption that their distribution will determine the distribution of the estimator as it moves from each point. The Central Limit Theorem tells us that, asymptotically, the distribution of these tangent vectors is normal with variance given by

$$\int_x \partial_i \ln p(x, \theta) \cdot \partial_j \ln p(x, \theta) \cdot p(x, \theta_0) dx$$

where θ is any point on the manifold where the estimate lies and θ_0 is the true parameter. This is of course the g^{θ_0} metric. Thus the limiting Brownian motion (maybe with drift) could be seen to be on a manifold with the metric given by the g^{θ_0} metric. Lemma 8.4.1 again takes care of the integrability condition for us just as in the previous case.

Again this is only a conjecture and this model needs careful study to see if there is any justification possible. The results showing high order efficiency in the curved exponential case do suggest that there might be some hope.

8.6. Preferred Point Geometry and Comparison of Estimators.

Working under the assumption that the model of section 8.3 is a good one for the understanding of the behaviour of the maximum likelihood estimator, we can apply some of the preferred point geometry theory which we developed in Chapter 7 to understand this model. In particular it might be possible to construct a similar type of model for estimators other than the maximum likelihood. If so the preferred point geometry can give us new ways of comparing estimators apart from the traditional one of efficiency.

We recall that the construction of the preferred point geometry model of the behaviour of the m.l.e. depended on the assumption that we had a large sample size and also on the assumption that the small scale behaviour of the model was all that was important. Thus we could sum up the behaviour at each point by the metric in the tangent space at that point. If this is a useful thing to do then we could repeat the procedure with other estimators using their small scale behaviour to define a metric at each point. For example, we could use the

method of moments estimator and construct a geometry which models its behaviour. This would similarly be a preferred point geometry.

We can now see how our understanding of preferred point geometry can enable us to compare the behaviours of various estimators. Since we are dealing with preferred point geometries the exact geometry will of course depend on the exact value of the preferred point. Therefore any information which we get from the estimator (via our model) has to be understood in context of our assumption of the true parameter. A measure of this conditioning is what we defined to be the *uncertainty* of the preferred point system, (see Chapter 7 for details). We recall that in the geometry which we are using to model the behaviour of the m.l.e. we calculated the uncertainty and found the nice result that in the exponential family case, for example, the geometry had zero first order uncertainty. However the total uncertainty was certainly not zero. We can, therefore, put forward the following question. For particular examples can we find estimators whose modeling preferred point geometries have smaller uncertainty than the m.l.e.? Does the method of moments geometry have this property for example? Thus the preferred point geometry analysis of estimators raises the question of whether there exists an estimator which, while it might be less efficient than the m.l.e., might possibly have smaller uncertainty under certain circumstances. Hence, in some conditions it might be preferable. The circumstances which we have in mind here include the situation when there is a misspecification of the true parameter. Again as with all the other issues raised in this chapter this question has to remain open and requires further study.

8.7. Global Geometry and Singularities.

In Chapter 1 we stated that all our geometry was of a local nature. In doing this we were following the literature. This gives various possible paths for the generalisation of this work. The first possible path is while working locally we can relax the regularity conditions of Chapter 1 which ensure that we are always working on a regular p -dimensional manifold. We could extend our possible geometries to include manifolds with singularities. There has been some work in this direction already. We in particular note the well known result which equates a nonsingularity in the Fisher information matrix with problems with identification in the parameter space. We see that it is the non-singularity of the metric which is implied by the fact that we are working on a manifold without singularities, (see section 1.4). In the context of our preferred point manifolds we see that the question of this non-singularity of the manifold is a

Future Work

very natural one. We notice that whenever we proved that our preferred point manifolds were metrics then it was always a proof of non-singularity in some neighbourhood of the true parameter. Therefore non-singularity outside of this open set would easily fit into our structure. We should point out that the points of nonsingularity themselves depend on the value of the preferred point.

Another form of nonsingularity which could be studied is a more basic one which gets away from the singularity of the metric and looks at the local topological structure of the underlying geometric space. An example of this is a family of distributions given by the mixture of two identical one dimensional exponential families

$$p(x, \lambda, \theta, \psi) = \lambda \exp\{\theta \cdot x + a(\theta)\} + (1 - \lambda) \exp\{\psi \cdot x + a(\psi)\}$$

where $\theta, \psi \in \mathbb{R}$ and $\lambda \in [0, 1]$.

We can show, although we shall not do that here, that the underlying parameter space is not a manifold, but rather a manifold with singularities. In fact it is a copy of the half space $\{(x, y, z) \mid z > 0\} \subset \mathbb{R}^3$ with a copy of \mathbb{R} glued on. The line of singularities represents the points in the space where the identification problem caused by either $\theta = \psi$ or by $\lambda = 0$ or 1 causes the parametric family to cease to be three dimensional. It would be interesting to study other parametric spaces which contain singularities and classify these singularities. Then it might be possible to get a better understanding of the identification problem from this geometric viewpoint.

The second possible path which can be taken in the generalisation of the geometric work is, while staying on regular manifolds, to work globally rather than locally. Thus we can consider spaces which have a more complex topology than just Euclidean space. It would appear however that developments here would be dependent on the work on the role of geodesics in parametric space which we discussed in some of the earlier sections. Some of the questions which could then be asked would be: What would the non-uniqueness of geodesics, which a non-trivial topology often brings mean from a statistical point of view? Also what are the implications of the non-trivial topology for the existence of sufficient statistics and possible ancillary statistics?

8.8 Expected and Observed Geometries.

Another area which could be developed is the relationship between expected and observed geometries particularly in the preferred point context. Although Barndorff-Nielsen has produced observed versions of the α -connection it is not clear yet whether we can produce the corresponding observed preferred point structures. One possible approach is to view the observed metric as forming part of a metric type structure on the sample space rather than on the parameter space and then studying maps between this space and various preferred point geometries on the parameter space.

8.9 Numerical Work.

There is one line of approach which has to be tackled if this type of analysis is going to be properly useful in applications as well as in theoretical statistics. This is the actual solving of the differential equations which the work on differential geometry always throws up. In particular the calculation of geodesics and the related problem of constructing affine coordinates, where they exist, is very important. We are, therefore, interested in finding good algorithms and the writing of computer programs for solving these problems quickly.

Appendix

On the Differential Geometry of the Wald Test with Nonlinear Restrictions

This appendix is a copy of the joint paper [CM&S1]. We include it since it is an example of the ways that geometric analysis can be used in practical statistical problems. Also it is an illustration of the discussion in Chapter 1 on the role of geometry in statistics and also of the discussion in Chapter 8 on the role of geodesics in inference.

1: Introduction

Several papers (in particular Gregory and Veall (1985)(1986) and LaFontaine and White (1986)) have recently noted the existence of a serious difficulty in the application of the Wald test to nonlinear restrictions in finite samples. Essentially the problem lies in that a given hypothesis may be written in any number of ways which are algebraically equivalent under the null but differ nontrivially under any particular alternative. Since the Wald statistic is based on a first order Taylor series expansion of the function defining the null hypothesis the test statistic is not invariant to the particular algebraic form chosen to represent the null . Thus , in general, different algebraic expressions of precisely the same null hypothesis lead to different test statistics which have different rejection regions at the same asymptotic significance level. The corresponding tests therefore have different exact significance levels and different powers at the same alternative. Gregory and Veall (1986) conclude their Monte Carlo study of a particular nonlinear example by emphasizing "the need for an analytical resolution to the problem of Wald test sensitivity".

In this paper we attempt to provide such an analysis through the use of differential geometry which has recently found considerable application within mathematical statistics (see for instance the references in Barndorff Nielsen, Cox and Reid(1986) and Amari et al. (1987)). We have two distinct goals in this paper ; one is to provide a clear theoretical understanding of why the Wald statistic behaves as it does in the non-linear case and the second is to produce practical solutions to the problem. Both of these objectives are most easily achieved using the techniques and insight provided by differential geometry.

We start our analysis with a critical look at the assumptions and justification of the Wald test particularly in the non-linear context. We show how the Wald statistic corresponds to a hybrid geometric quantity, in that it considers a vector in a statistical manifold and yet measures its length using a metric which is only appropriate to a tangent space. We then follow this theoretical analysis by defining a truly geometric test in the correct space which is a direct generalisation of the Wald test to our non-linear case. This new geometric test we call *The Geodesic test*. We can therefore reach a theoretical resolution of the problems of the Wald statistic by showing how the dependence of the statistic on the form of the restriction function is viewed geometrically as a failure of the statistic to transform correctly under a change of coordinates of the underlying manifold. The geodesic test is shown to transform correctly and we can further show that the two tests coincide under the classical assumptions of the General Linear Model with linear restrictions and in this case the Wald test is reliable. The two tests are in any case asymptotically equivalent as we show below. We continue the geometric approach by using the tools of curvature and the related notion of the Christoffel symbols of a metric to compare the two statistics in general .

As a first practical result we show how these Christoffel symbols can be used to compare different forms of the restriction function and hence how a best selection can

be made. As an example of these methods we consider the discrimination between the two choices of restriction function used by Gregory and Veall (1986).

We continue by studying how the choice of restriction function effects the behaviour of the Wald statistic using graphical techniques suggested by the previous analysis. Finally we establish, under certain regularity conditions an inequality between the Geodesic and Wald statistics that indicates when reliable inference is possible using the Wald test. This last practical application of our geometric approach completes the paper.

Three papers, in particular, have appeared recently that are related to our analysis. Moolgavkar and Venson (1987), have employed a geometrical analysis of Wald confidence intervals for a simple hypothesis in nonlinear regression and more generally curved exponential families. Their object is to reparametrise the model so that it looks as much like " uncurved" Euclidean space as possible. One way to achieve this is to use geodesic normal coordinates but as we shall see below, given the difficulty of calculating geodesics in practise they are forced to use approximations that while they improve on the Wald confidence regions they do not correspond with the geodesic regions that follow from the geodesic statistic we introduce below.

Veath (1985) also considers the use of reparametrisation in the exponential model , however the restriction of his analysis to the one dimensional case avoids much of the difficulty of the multidimensional problem that we consider below. The results of both these papers are encompassed by in our more general geometric procedures below.

Phillips and Park (1988) have also considered the issue by means of calculating Edgeworth expansions to investigate alternative forms of the Wald statistic with nonlinear restrictions. These expansions are able to explain , to a degree ,the observed behaviour of the test as the higher order terms account for the deviations from the asymptotic distribution and also to provide corrections to the test that indicate transformations of the restrictions which accelerate convergence to the asymptotic distribution. However the analysis is limited to the $O(T^{-1})$ terms in the expansions and hence their correction factors are similarly limited unlike the geometric analysis and Geodesic test introduced below.

2: The Wald Test

The algebraic development of the Wald statistic may be found in any standard text such as Silvey (1975) or Cox and Hinkley(1973) and assumes a model summarised in a log likelihood function $l(.,\theta)$ together with an estimator $\hat{\theta}$ for the unknown parameter $\theta \in \mathbb{R}^p$, which is distributed at least asymptotically as multivariate normal $N_p(\theta, I_\theta^{-1})$. This happens of course in all regular likelihood problems where we can

identify I_θ with Fisher's information matrix. The null hypothesis is specified as the zero level set of a vector valued function g :

$$H_0 = g^{-1}(0) = \{\theta \in \Theta \mid g(\theta) = 0\} \quad (1)$$

where Θ denotes the parameter space and $g = (g_1, \dots, g_r)$ is a vector of real valued functions, one for each restriction. The Wald statistic, $W(g)$, is then defined as

$$W(g) = g(\hat{\theta})^T \{\text{var } g(\hat{\theta})\}^{-1} g(\hat{\theta}) \quad (2)$$

in which the estimated variance covariance matrix of $g(\hat{\theta})$ is given by

$$\text{var } g(\hat{\theta}) = [Dg(\hat{\theta})]^T I_{\hat{\theta}}^{-1} [Dg(\hat{\theta})] \quad (3)$$

where $Dg(\hat{\theta})$ is the evaluation at $\theta = \hat{\theta}$ of the pxr matrix

$$Dg(\theta) = \left(\frac{\partial g_i(\theta)}{\partial \theta_j} \right) \quad (4)$$

and $I_{\hat{\theta}}$ is the evaluation of I_θ at $\theta = \hat{\theta}$.

Since $W(g)$ depends solely upon quantities evaluated at $\hat{\theta}$, it is particularly well suited for use in situations where the unrestricted estimate $\hat{\theta}$ is easy to compute but the restricted maximum likelihood estimate, $\tilde{\theta}$, under H_0 is not. This is likely to be the case when the restriction function $g(\theta)$ is nonlinear and yet it is precisely in this case that the difficulties with the use of the Wald test appear.

The distribution and properties of the Wald Statistic, which is based on an expansion of the restriction function $g(\theta)$, rest on three fundamental approximations:

(i). Ignore any non-normality in the finite sample distribution of $\hat{\theta}$, in other words work effectively only with the asymptotic distribution,

$$\sqrt{n}(\hat{\theta} - \theta) \sim N_p(0, B_{\theta_0}^{-1}),$$

where B_{θ_0} is the information matrix for a single observation, and θ_0 is the assumed true value of θ .

(ii). Ignore all terms beyond the linear one in the Taylor expansion of $g(\theta)$ about θ_0 evaluated at $\hat{\theta}$.

$$\sqrt{n}\{g(\hat{\theta}) - g(\theta_0)\} = \sqrt{n}[Dg(\theta_0)]^T(\hat{\theta} - \theta_0) + O(\sqrt{n} \mid \hat{\theta} - \theta_0 \mid^2)$$

in other words work effectively with

$$\sqrt{n}\{g(\hat{\theta})-g(\theta_0)\} \sim N_r(0, [Dg(\theta_0)]^T B_{\theta_0}^{-1} [Dg(\theta_0)])$$

- (iii). Finally to gain an operational statistic, ignore the dependence in the covariance matrix of $(g(\hat{\theta})-g(\theta_0))$ on the unknown θ_0 and replace θ_0 by $\hat{\theta}$. In other words use

$$\sqrt{n}\{g(\hat{\theta})-g(\theta)\} \sim N_r(0, [Dg(\hat{\theta})]^T \hat{B}_{\hat{\theta}}^{-1} [Dg(\hat{\theta})])$$

Under these conditions $W(g)$ is asymptotically a χ_r^2 random variable under H_0 .

The applicability of the Wald statistic is critically determined by the validity of these approximations. In particular (i) covers any regular maximum likelihood problem and those where a central limit theorem may be applied. Approximation (ii) is exact only if $g(\theta)$ is affine i.e. $g(\theta)=A\theta +b$, and (iii) is exact only if $g(\theta)$ is affine and I_θ is independent of θ . This latter condition we refer to as the constant metric case below. Critically for our present concern it is the linearization in (ii) that leads to the lack of invariance with respect to reparameterizations.

The previous argument leads to the standard Wald statistic as implemented empirically. However for our geometric analysis of the statistic we abstract from the final approximation which replaces the unknown θ_0 by the observed $\hat{\theta}$. While clearly necessary for practical implementation of the statistic this final step introduces unnecessary elements and complexities for our theoretical analysis. The source of the problems with the Wald statistic with which we are concerned lies in the first two approximations, hence we shall consider below a Wald statistic of the form

$$W_{\theta_0} = g(\hat{\theta})[Dg(\theta_0)^T I_{\theta_0}^{-1} Dg(\theta_0)]^{-1} g(\hat{\theta}) \quad (5)$$

rather than the usual $W_{\hat{\theta}}$ which is defined as we have already stated as

$$W_{\hat{\theta}} = g(\hat{\theta})[Dg(\hat{\theta})^T I_{\hat{\theta}}^{-1} Dg(\hat{\theta})]^{-1} g(\hat{\theta}) \quad (6)$$

where the covariance matrix of $g(\hat{\theta})$ is evaluated at $\hat{\theta}$. Both of the statistics (5) and (6) imply a fixed metric and having conducted our theoretical argument in terms of statistic (5) it can be easily shown that precisely the same implications apply to the empirical Wald statistic (6). Our recommendations for a practical solution to the lack of invariance of the Wald Statistic apply in particular to the applied statistic (6).

3: The geometry of the Wald Statistic and the Geodesic Test.

3.1 An overview.

As discussed above the construction of a Wald statistic enables a standard test for hypotheses expressed on some parameter space indexing a family of distributions to be performed. If we parameterise this space, Θ , by $(\theta_1, \dots, \theta_p)$ then at least locally we may, without loss of generality, write the null hypothesis as the zero set of the restriction function, g , where,

$$g: \Theta \rightarrow \mathbb{R}^r$$

is a smooth enough function. So the null hypothesis is the subset of Θ given by

$$H_0 = \{(\theta_1, \dots, \theta_p) \mid g(\theta_1, \dots, \theta_p) = 0\}$$

For the generality we need in our analysis we take the space of distributions to be nonlinear or curved although even when it may be linear, as we shall see below, the effect of a nonlinear restriction is to introduce nonlinearity to the structure of the space.

The Wald test attempts to measure the probability of deviations from the null by constructing contours, using the mathematical form of the statistic, around the null hypothesis. The Wald statistic then takes positive values as the estimated value of the parameter lies outside some chosen contour.

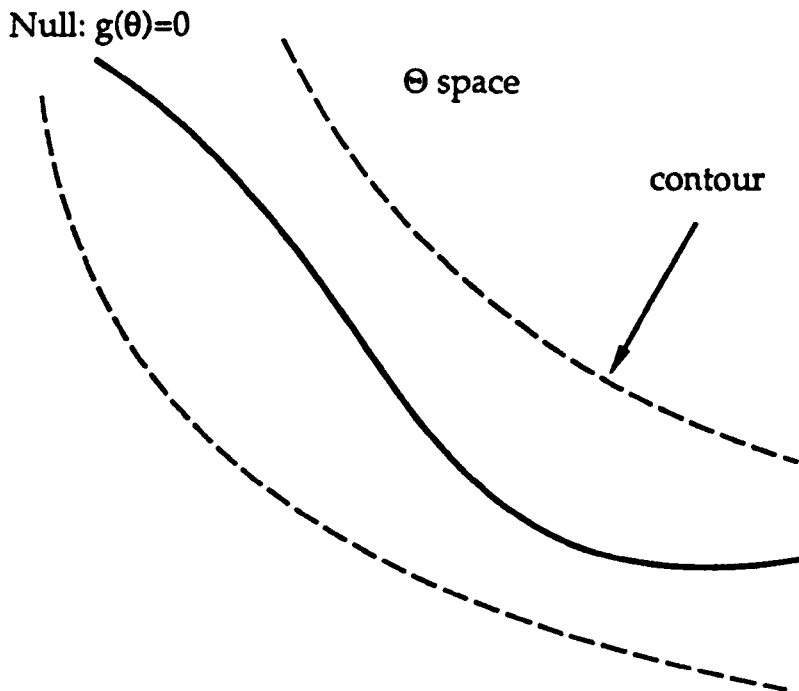


Fig.1

The problem with the Wald Test is that although many functions can equally be used to define the null hypothesis H_0 the approximation that is used to estimate the probability contours depends on which particular function is used. The reason loosely for this is although two different functions $g_1, g_2: \Theta \rightarrow \mathbb{R}^r$ may agree on their zero set, i.e.

$$\{(g_1)^{-1}(0)\} = \{(g_2)^{-1}(0)\},$$

the sets that correspond to their other levels, $\{(g_1)^{-1}(c)\}$ and $\{(g_2)^{-1}(c)\}$ ($c \neq 0$), can differ arbitrarily. Moreover unlike the one dimensional case, considered by Væth (1985), this problem cannot be resolved by a simple rescaling.

This implies that the reason for the inconsistency in the performance of the Wald test stem from this difference in behaviour of the level sets away from the null hypothesis which is not taken into account by the Wald statistic.

Geometrically we can see, given approximation (ii) above, that the Wald statistic has an interpretation as the squared length of a particular vector valued function, $(g(\hat{\theta}) - 0)$, on the curved manifold describing the family of potential distributions. The metric used to calculate the length of this vector is however taken, as will be shown below, from the tangent space to the manifold at θ_0 . The Wald statistic for a nonlinear restriction is therefore a hybrid quantity measuring a vector corresponding to a point in a nonlinear, non-metric, space (the statistical manifold) with a metric taken from a linear tangent space. Notice that the Likelihood Ratio and Lagrange Multiplier statistics do not suffer from this inconsistency. The Likelihood Ratio statistic being simply a comparison of values taken by the likelihood function on the manifold and the Lagrange multiplier measuring the length of a vector in the tangent space with a consistent metric.

When the natural coordinate system defined by Θ is employed the lengths of tangent vectors are measured using the Fisher information metric, I_θ , at each point. The following diagram demonstrates the situation in general. The curved manifold corresponds to the nonlinear space of distributions indexed by the choice of θ or alternatively the value of $g(\theta)$ and for each point on this manifold there will be a tangent plane on which its associated metric is defined.

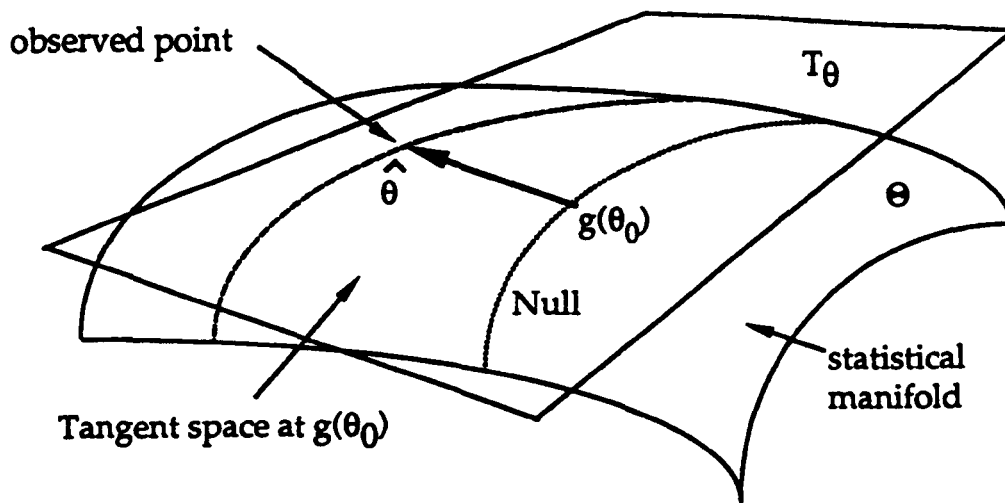


Fig. 2

Notice also that since we are free to choose any coordinate system to describe the statistical manifold and since the Wald statistic is itself indexed by the choice of restriction function $g(\theta)$ it is necessary to consider the manifold in terms of a new coordinate system that involves $g(\theta)$ rather than the natural coordinates $\{\theta_i\}$. This change in coordinate system however also induces a change in form of the metric used to measure distance in the tangent space which the Wald statistic exploits. Choosing a particular algebraic function, g_1 , to express the restriction in effect then imposes a choice of (probably non-Euclidean) coordinates on the statistical manifold. Choosing a different function, say g_2 , therefore induces a change of coordinates, or a reparametrisation of the manifold and changes the form of the metric.

From this geometric point of view it is then possible to see how the Wald statistic does not transform in the correct way under the change of coordinates which corresponds to a different choice of restriction function, thus causing the inconsistencies in its behaviour. The Wald statistic is essentially a quadratic form on a linear space, the tangent space, which is appropriate to measuring the length of (linear) vectors in this space. The statistic transforms appropriately for linear transformations but inadequately for nonlinear transformations induced by nonlinear restrictions. In addition any nonlinear coordinate system on the manifold implies a different metric will be appropriate for every point on the manifold while the Wald statistic implicitly assumes that there is a single metric for the entire manifold. It is only under this constant metric assumption that the Wald statistic provides a well defined measure of length. Given this geometric insight we can see that Wald statistics computed for different nonlinear restriction functions are not comparable.

There are two main reasons why nonconstant metrics may come about in general. The first is that the underlying manifold has non-zero curvature and so there simply is no coordinate system which would give a constant metric. The variation of the metric may also be induced by the particular choice of coordinate

system . This distinction corresponds with that made earlier by Bates and Watts (1980) between "intrinsic" and "parameter effects" curvature, where intrinsic curvature cannot be removed by a reparametrisation of the problem. Notice that even for a space with no curvature at all most coordinate systems will not give the constant representation of the metric that the Wald statistic requires. The property of a constant metric representation leads to what is known as an *Affine coordinate system*. An example of a non-affine coordinate system on a flat space is the use of polar coordinates $(r\cos\psi, r\sin\psi)$ on the Euclidean plane. Here the standard metric will be given by the form.

$$\begin{bmatrix} r & 0 \\ 0 & 1 \end{bmatrix}$$

which is a nonconstant metric since it depends on the point (r, ψ) .

In what follows we derive an alternative approach for calculating confidence regions in a space with such a varying metric. The resulting test statistic , the Geodesic statistic, has the advantage of behaving properly under changes of coordinates, and hence different choices of restriction function. This Geodesic statistic has a geometric interpretation as the length of a curve in the statistical manifold itself rather than in the tangent space and is invariant under coordinate transformation. We discuss cases in which it reduces to the standard Wald statistic and hence when the use of the Wald test will be free of its dependence on the form of restriction function.

Figure 3 below demonstrates our strategy. Initially the statistical problem is formulated in the $\{\theta_i\}$ coordinate system, and we change coordinates to a new coordinate system, (g, k) , where g is the value of the restriction function and the remaining coordinates, k , are, without loss of generality, chosen orthogonally to g . Working in this particular , (g, k) , coordinate system we can clearly see the geometric interpretation for the Wald Statistic as the length of some vector in a tangent space. Ideally this vector would correspond to the correct projection from the manifold to the tangent space of the point $\hat{\theta}$, the unrestricted parameter estimate. This projection is achieved by what is known as the exponential map which preserves the correct length of the implied vector. We use this map to show how the Wald statistic does not transform properly with respect to changes in coordinates.

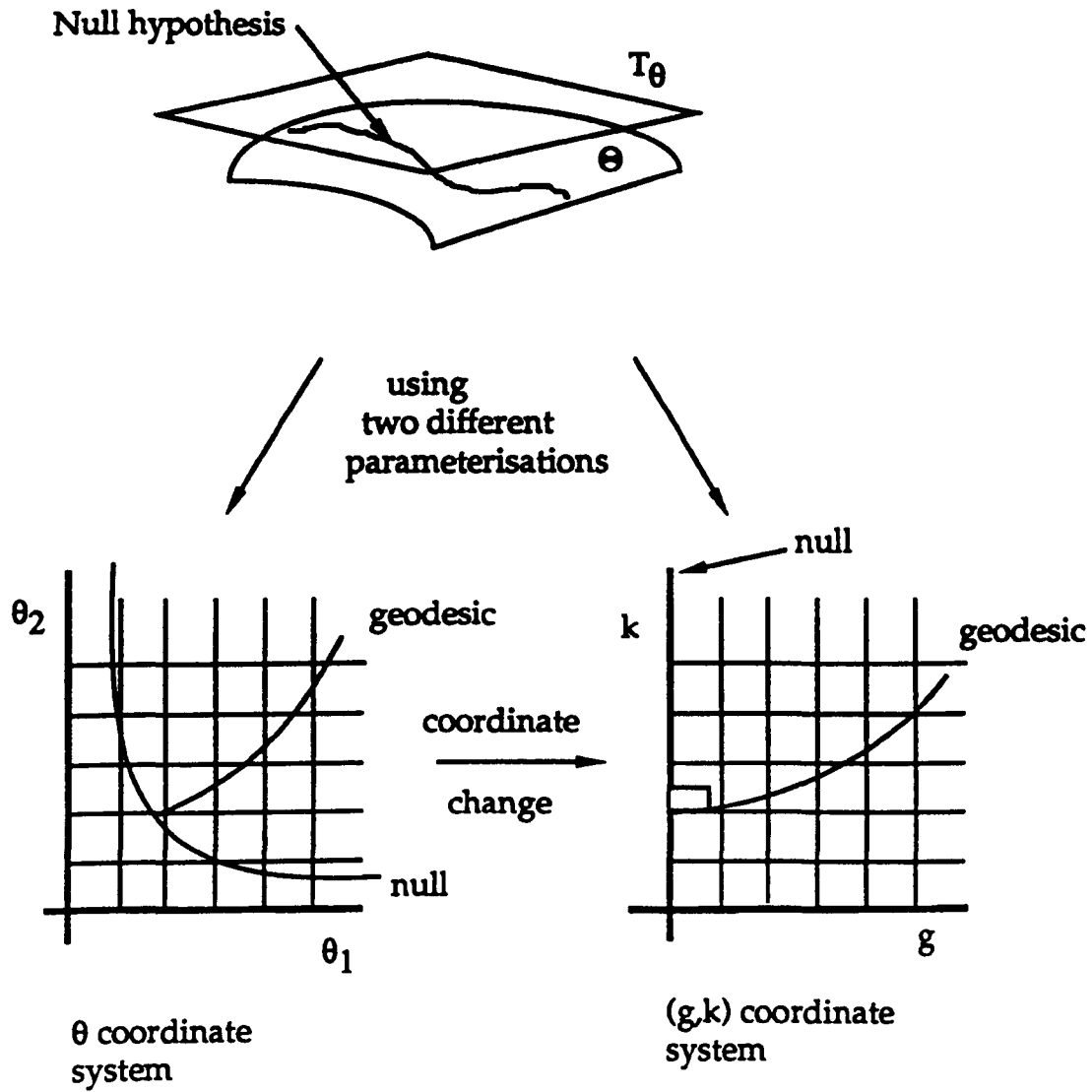


Fig. 3

3.2: Formal Analysis

The following lemma shows how the choice of restriction function , g , determines a choice of coordinates.

Lemma 3.2.1:

- (a) If θ_0 is any point in Θ such that $Dg(\theta_0)$ has full rank, then in an open neighbourhood of θ_0 there exists a local coordinate system of the form,

$$(g_1(\theta), \dots, g_r(\theta), k_1(\theta), \dots, k_{p-r}(\theta))$$

Appendix

where $g(\theta) = (g_1(\theta), \dots, g_r(\theta))$ represents the r -vector of restrictions and k_1, k_2, \dots, k_{p-r} are real valued smooth functions on R^p . Furthermore if $F(,)$ represents the Fisher information metric, then at any point on the level set of g to which θ_0 belongs we have that;

$$F\left(\frac{\partial}{\partial g_i}, \frac{\partial}{\partial k_j}\right) = 0 \quad \forall i, j \quad (7)$$

where $\frac{\partial}{\partial g}$ and $\frac{\partial}{\partial k}$ form a basis for the coordinate system where the vector $k(\theta)$ is chosen such that the corresponding tangent vectors satisfies (7), as shown in the following diagram.

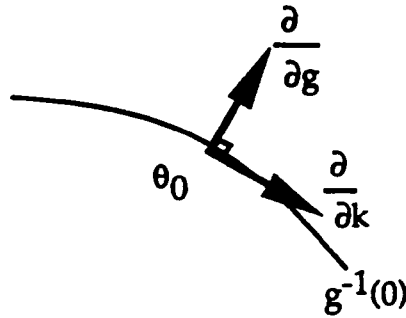


fig. 4

- (b) In the case where the null hypothesis is given by just a single restriction we have that the above orthogonality condition is true for any point in an open region around the level set of $g(\cdot)$ to which θ_0 belongs and not just on the one level set.

Proof: All proofs are given in the appendix to the paper.

The implication of this lemma lies in that we must consider how the Fisher information metric is transformed as the coordinate system in which the statistical hypothesis is expressed itself varies with the change of form of g . Let I_θ be the Fisher information matrix and $(\theta_1, \dots, \theta_p)$ our original coordinate system on Θ then if we let $G = (\partial g_i / \partial \theta_j)$ we have the following result.

Lemma 3.2.2:

(a) In terms of the (g,k) coordinate system given above the matrix defining the Fisher Information metric at a point θ is given by

$$F_g = \left(\begin{pmatrix} G \\ K \end{pmatrix}^{-1} \right)^T I_\theta \begin{pmatrix} G \\ K \end{pmatrix}^{-1} \quad (8)$$

where $K = (\partial k_i / \partial \theta_j)$ is the $(p-r) \times p$ matrix that induces the change in coordinates for vectors parallel to the level sets of g , i.e. the vectors $\left\{ \frac{\partial}{\partial k} \right\}$.

(b) In the single restriction case the formula holds in an open region of the g level set of θ_0 .

Corollary 3.2.3:

If X is a vector field always orthogonal to the level sets of g , then working in our (g,k) coordinate system we see that the squared length of X at all points θ is given by $X^T (G^T I_\theta^{-1} G)^{-1} X$. This reduces to the either of our two forms of the Wald test statistic (5) or (6), depending on where G and I_θ are evaluated given that X equals $g(\hat{\theta})$.

Notice in fact that the Wald test considers the length of the vector $(g_1, \dots, g_r, 0, \dots, 0)$ which is orthogonal to the level sets of g and hence lies in the vector field X which is defined above. Any vector in X then has its length measured by the formula given in the corollary. This corollary shows the difficulty with the use of the Wald test lies in that instead of being a measure of a length in the curved manifold it is in fact measuring a length in the flat tangent space. It is this confusion between the manifold and its tangent space at a point which is causing the statistic to be dependent on the choice of coordinate system, and through the coordinate system the statistic ultimately depends on the particular algebraic form of the restriction. The difference between the two spaces is that while on the manifold the form of the Fisher metric changes from point to point, the tangent space is a linear space with a constant metric.

Appendix

To understand the relationship between these two geometric structures we need to introduce the notion of the exponential map between the tangent space at a point θ and the manifold.

$$\exp: T_\theta \rightarrow M.$$

This exponential map is defined in the following way; $\exp_\theta(v)$ is the point which lies on the geodesic starting at θ in the direction v which is a geodesic distance $|v|$ from θ .

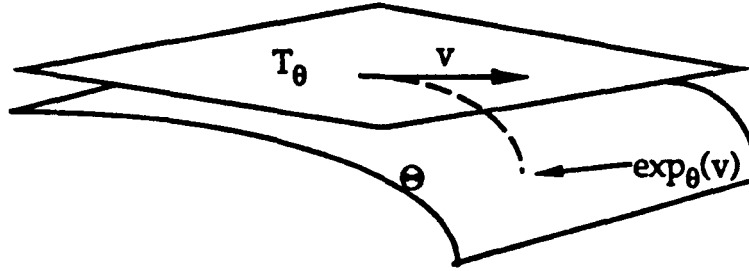


Fig. 5

Where,

Definition: Path length on the manifold

If $\gamma(t):[0,1] \rightarrow \Theta$ is the path starting at θ_1 and ending at θ_2 then the path length is given by,

$$\int_0^1 \sqrt{H\left(\frac{d}{dt}\gamma(t), \frac{d}{dt}\gamma(t)\right)} dt \quad (9)$$

where H is the metric .

Definition: A geodesic in the manifold with a metric is a curve $\gamma(s)$ which has the shortest path length between two points, where s is the arc length parameter. It can be characterised locally in our manifold as being the solution to the set of second degree differential equations given by:

$$\ddot{\gamma}^i + \Gamma_{jk}^i \dot{\gamma}^j \dot{\gamma}^k = 0 \quad i, j, k = 1, \dots, p \quad (10)$$

where Γ_{jk}^i are the Christoffel symbols for the Levi-Civita connection (see for instance (Amari 1985)). These symbols are determined from the metric (h_{ij}) , and its inverse (h^{ij}) , by the equations:

$$\Gamma_{jk}^i = \frac{1}{2} h^{mi} \left(\frac{\partial h_{mk}}{\partial \theta_j} + \frac{\partial h_{jm}}{\partial \theta_k} - \frac{\partial h_{jk}}{\partial \theta_m} \right) \quad (11)$$

Applying these definitions to our problem we have,

Definition: Geodesic distance between θ and $\hat{\theta}$

If $\gamma(t):[0,1] \rightarrow \Theta$ is the geodesic starting at $\theta \in g^{-1}(0)$ and ending at $\hat{\theta}$ then the geodesic distance $(\theta, \hat{\theta})$ is given by,

$$\int_0^1 \sqrt{F_g\left(\frac{d}{dt}\gamma(t), \frac{d}{dt}\gamma(t)\right)} dt \quad (12)$$

where F_g is the Fisher information metric and t is a parameter on the geodesic which is zero on the null hypothesis and 1 at the observed point in the manifold.

The Wald statistic defined in general as

$$(g(\hat{\theta}) - g(\theta))^T (G_\theta^T I_\theta^{-1} G_\theta)^{-1} (g(\hat{\theta}) - g(\theta))$$

then represents the squared length of $(g(\hat{\theta}) - g(\theta))$ measured in the tangent space at a point θ . Whereas from the geometric point of view we would ideally wish to identify a point in the manifold corresponding to $\exp(g(\hat{\theta}) - g(\theta))$. Notice in order to measure the length of $\exp(g(\hat{\theta}) - g(\theta))$ we need to consider the sequence of metrics corresponding to the sequence of tangent planes which are based on those points on the geodesic line to $\exp(g(\hat{\theta}) - g(\theta))$ from θ . The Wald statistic however is defined at the one tangent space based at θ and measures the corresponding length using the fixed metric from that one tangent space. The importance of this distinction arises when there is a change in coordinates such as that induced by a different choice of restriction function. The exponential map remains unaffected by this transformation since it considers the change of basis at each tangent space on the manifold. The Wald statistic is only determined by the change of basis in the one tangent space based at θ . Moreover the Wald statistic measures the length of a different vector and will not be comparable with that calculated using the original choice of restriction function.

Following this logic we are naturally led to introduce as an alternative to the Wald statistic the Geodesic statistic which follows the standard differential geometric construction for measuring the distance between two points on a manifold with a metric, a geodesic. As discussed above this geodesic statistic has the advantage of being coordinate free thus escaping the problem of being dependent on the choice of restriction function.

By changing the parametrisation on the geodesic to the value of the function g at each point we see that (12) is equivalent to

$$G = \int_0^{g(\hat{\theta})} \sqrt{F_g\left(\frac{d}{dg}\gamma(g), \frac{d}{dg}\gamma(g)\right)} dg \quad (13)$$

Where $\frac{d}{dg}\gamma(g)$ is the tangent field along the geodesic $\gamma(g)$ which we are using to measure the distance from the point θ .

We can therefore now formally define the Geodesic Statistic.

Definition: The geodesic statistic. For any point in our manifold, corresponding to $\hat{\theta}$ we can measure the geodesic distance (13) to any point on the null hypothesis. The minimum such length provides the value of the geodesic statistic.

The finite sample distribution of this statistic is uncertain but is considered in more detail in Critchley, Marriott and Salmon (1989b). However asymptotically at least we are assured that this statistic will be distributed χ^2_r and this forms the basis for the proposed Geodesic Test. As we show below this limiting distribution follows since asymptotically as $\hat{\theta}$ converges to the null the distinction between the Wald and Geodesic statistics vanishes and indeed the Wald statistic itself becomes immune to the problems of reparametrisation in this nonlinear environment. Although as shown by Philips and Park the speed of convergence to the limiting distribution may be critically determined by the choice of restriction function. These conclusions follow from the fundamental property of the geodesic which generates the geodesic statistic which is that it starts perpendicular to the null hypothesis before reaching the point $\hat{\theta}$.

One essential difference between the two statistics in finite samples lies in that the Wald statistic ignores the component of the total information held in the k -coordinates whereas the geodesic statistic exploits this ancillary information. More generally the following lemma establishes the conditions under which Wald and geodesic inference will coincide.

Lemma 3.2.4:

In the single restriction case the Wald test statistic will agree with the squared geodesic distance if

- (i) F_g , the matrix representation of the metric is constant throughout the manifold,
- (ii) the geodesics between any $\hat{\theta}$ and the null hypothesis are perpendicular to the level sets of g .

These conditions hold if we are working in Euclidean space and our restriction is just a linear function (as in the general linear model) so that all the level sets are parallel lines and all the geodesics are just orthogonal straight lines. Note that because the restriction function is linear the metric will stay constant in the (g,k) coordinate system. It is the second condition in this lemma that eliminates the dependence of the information in the k -coordinates.

4: An Error bound for the Wald statistic

In general it may be hard to calculate the statistic required for the geodesic test since it requires the solution to a set of second order quasi-linear differential equations (10), and then integrating along these curves. Both of these operations are , in general, difficult analytically although numerical methods are available to provide approximate solutions for a given example (Marriott and Salmon (1989)). In the general linear model with a constant metric in the natural coordinate system and nonlinear restrictions it is however possible to find explicit solutions to the geodesics which may then be evaluated numerically . An alternative and completely general approach that we follow in this section of the paper is to calculate a bound between the Wald and geodesic statistics . In this way we are able to determine whether the Wald statistic seriously deviates for a given form of restriction function and in addition we obtain a formal basis to compare different forms of restriction function. Thus we wish to consider the two statistics given by (13) and

$$W_{\theta_0} = g(\hat{\theta})[Dg(\theta_0)^T I_{\theta_0}^{-1} Dg(\theta_0)]^{-1} g(\hat{\theta})$$

The difference between these two statistics can be measured by

$$|G - \sqrt{W}| = \left| \int_0^{g(\hat{\theta})} \sqrt{F_g\left(\frac{d}{dg}\gamma(g), \frac{d}{dg}\gamma(g)\right)} dg - \sqrt{g(\hat{\theta})^T (Dg(\theta_0)^T I_{\theta_0}^{-1} Dg(\theta_0))^{-1} g(\hat{\theta})} \right| \quad (14)$$

where the two statistics are essentially of the same form , representing line integrals in the manifold expressed in the (g,k) coordinate system, except that the Wald Statistic reduces to its simpler form through its use of a constant metric at θ_0 and its independence of any information in the orthogonal direction given by the k coordinates. Hence the difference may be rewritten as

$$|G - \sqrt{W}| = \left| \int_0^{g(\hat{\theta})} \sqrt{F_g\left(\frac{d}{dg}\gamma(g), \frac{d}{dg}\gamma(g)\right)} dg - \int_0^{g(\hat{\theta})} \sqrt{F_{g=g(\theta_0)}\left(\frac{d}{dg}\gamma(g), \frac{d}{dg}\gamma(g)\right)} dg \right| \quad (15)$$

By applying the mean value theorem for differentiable functions we see that;

$$|G - \sqrt{W}| \leq \left| \int_0^{g(\hat{\theta})} \max_g \frac{d}{dg} \left\{ \sqrt{F_g\left(\frac{d}{dg}\gamma(g), \frac{d}{dg}\gamma(g)\right)} \right\} dg \right| = g(\hat{\theta}) \max_g \left\{ \frac{d}{dg} \sqrt{F_g\left(\frac{d}{dg}\gamma(g), \frac{d}{dg}\gamma(g)\right)} \right\} \quad (16)$$

Hence we can see that the two statistics will be the same if $F_g(,)$ is constant for all values of g and so a choice of restriction function that induces the least possible variation in the metric is clearly seen to be preferred.

To make any further comparison we need to consider how the metric F_g varies with g and the rest of this section is devoted to providing an explicit bound on the difference between the two statistics based on the difference between the square of the geodesic length and the Wald statistic since the Wald statistic in fact represents a squared distance measure. For clarity we restrict our attention to the two dimensional case where Θ is a surface and the null hypothesis a curve although the analysis may easily be extended to higher dimensional manifolds, in particular if there is only one restriction function this is particularly easy. Working in the (g,k) coordinate system the Fisher metric is given by the matrix

$$\begin{bmatrix} f_{11} & 0 \\ 0 & f_{22} \end{bmatrix}$$

where by lemma 3.2.1(b) $f_{12}=f_{21}=F(\partial/\partial g, \partial/\partial k)=0$ by definition of the (g,k) coordinate system. The geodesic, $\gamma(g)$, is a curve parametrised by the value of the function g and therefore in these coordinates may be written $(g, \phi(g))$, hence we have that

$$F_g\left(\frac{d}{dg}\gamma(g), \frac{d}{dg}\gamma(g)\right) = (f_{11} + (\phi')^2 f_{22}). \quad (17)$$

Considering this expression in more detail it is clear that we need to understand both how the geodesic behaves with g and also how the form of the metric itself varies with g . We now take each of these questions in turn.

For a general analysis, without an explicit form for the geodesic, we are forced to consider its dependence on g using the projection of the geodesic on \mathbb{R}^2 Euclidean space. This projection is defined by the (g,k) coordinate system as shown in the diagram below. Notice that although the geodesic itself will have zero curvature in the manifold its image will have nonzero curvature and we can use this fact to establish a bound on the behaviour of the geodesic. In addition the coordinates of the geodesic and its image coincide although the relevant metric in each case will differ. We start by considering the curvature of the image of the geodesic using the angle ω as shown in the diagram.

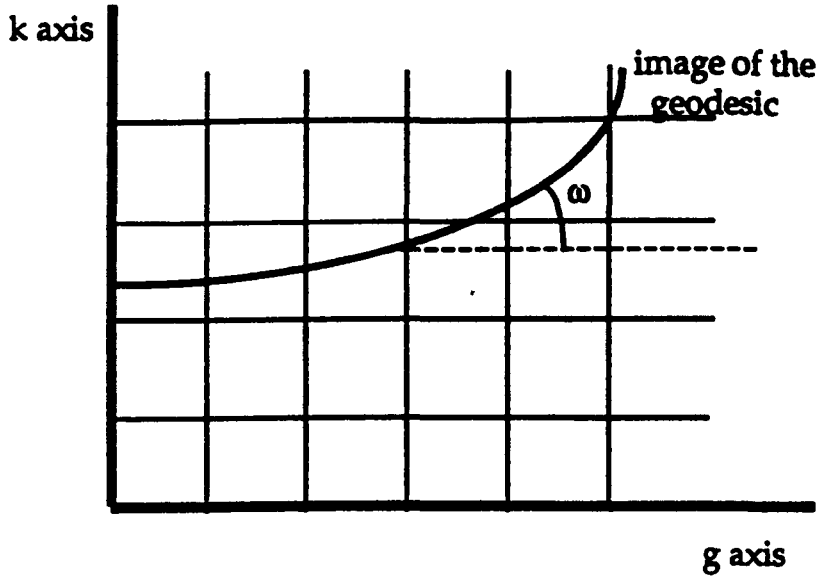


Fig 6.

We estimate the angle ω in the standard manner ; parameterising the image of the geodesic now by $(\alpha(s), \beta(s))$, where s measures arc length in the manifold, the curvature is given by,

$$\kappa = \frac{\ddot{\alpha}\dot{\beta} - \dot{\alpha}\ddot{\beta}}{(\dot{\alpha}^2 + \dot{\beta}^2)^{\frac{3}{2}}} \quad (18)$$

Since $\gamma(s) = (\alpha(s), \beta(s))$ is a geodesic of the surface it will satisfy by definition the differential equations (10) given earlier as

$$\ddot{\gamma}^i + \Gamma_{jk}^i \dot{\gamma}^j \dot{\gamma}^k = 0 \quad i, j, k = 1, 2$$

where Γ_{jk}^i are , as above, the Christoffel symbols for the Levi-Civita connection of the Fisher metric, with respect to the (g, k) coordinate system. We show how to calculate these symbols for a specific example later in the paper, but for the moment it is sufficient to know that they are determined entirely by the Fisher information metric . It should be noted that the Christoffel symbols are not themselves geometric objects and depend on the choice of coordinate system.

In order to calculate ω we write $\dot{\gamma}(s)$ as $(r \cos \omega, r \sin \omega)$ and hence we find the differential equations (10) defining the geodesic to be,

$$\begin{aligned}\tilde{\alpha} &= r^2(\cos \omega, \sin \omega) \begin{bmatrix} \Gamma_{11}^1 & \Gamma_{12}^1 \\ \Gamma_{21}^1 & \Gamma_{22}^1 \end{bmatrix} \begin{pmatrix} \cos \omega \\ \sin \omega \end{pmatrix} \\ \tilde{\beta} &= r^2(\cos \omega, \sin \omega) \begin{bmatrix} \Gamma_{11}^2 & \Gamma_{12}^2 \\ \Gamma_{21}^2 & \Gamma_{22}^2 \end{bmatrix} \begin{pmatrix} \cos \omega \\ \sin \omega \end{pmatrix}\end{aligned}\quad (19)$$

Hence the curvature of the image of the geodesic given by (18) is found to be

$$\kappa = \langle (\tilde{\alpha}, \tilde{\beta}), (-\sin \omega, \cos \omega) \rangle = -\cos \omega \frac{\ddot{\beta}}{r^2} + \sin \omega \frac{\ddot{\alpha}}{r^2} \quad (20)$$

which with the explicit values of $\tilde{\alpha}$ and $\tilde{\beta}$ given in (19) substituted gives κ as a function of the angle ω and the Christoffel symbols, (\langle, \rangle denotes the normal Euclidean product).

We are now in a position to consider how the geodesic varies with the choice of g , and this can be achieved even without an explicit form for the geodesic by considering bounds on the curvature of its image. A number of different bounds may be constructed and we present one simple and intuitive choice, others may in fact be tighter. The bound we propose exploits the maximum curvature of the image of the geodesic, however all such bounds are essentially determined by the values of the Christoffel symbols and so we show how these may usefully be regarded as criteria on which to base judgement about the particular choice of g function and its associated Wald test.

If we let λ_1 , and λ_2 denote the maximum (in modulus) of the eigenvalues of the matrices

$$\Gamma^i = \begin{pmatrix} \Gamma_{11}^i & \Gamma_{12}^i \\ \Gamma_{21}^i & \Gamma_{22}^i \end{pmatrix} \quad i=1,2 \quad (21)$$

taken over the relevant region of Θ space. Then equations (19) and (20) give the maximum curvature κ_{\max} , as

$$|\kappa_{\max}| \leq ((\lambda_1)^2 + (\lambda_2)^2)^{1/2}. \quad (22)$$

We can now use κ_{\max} to get an upper bound on ω and hence bound the behaviour of the geodesic. Figure 7 shows the situation, the angle ω which the tangent to the geodesic makes with the g -axis will be less than the angle made by any curve whose curvature is everywhere greater than that of the image of the geodesic. In particular the circle of radius $1/(\kappa_{\max})$, whose curvature is everywhere κ_{\max} .

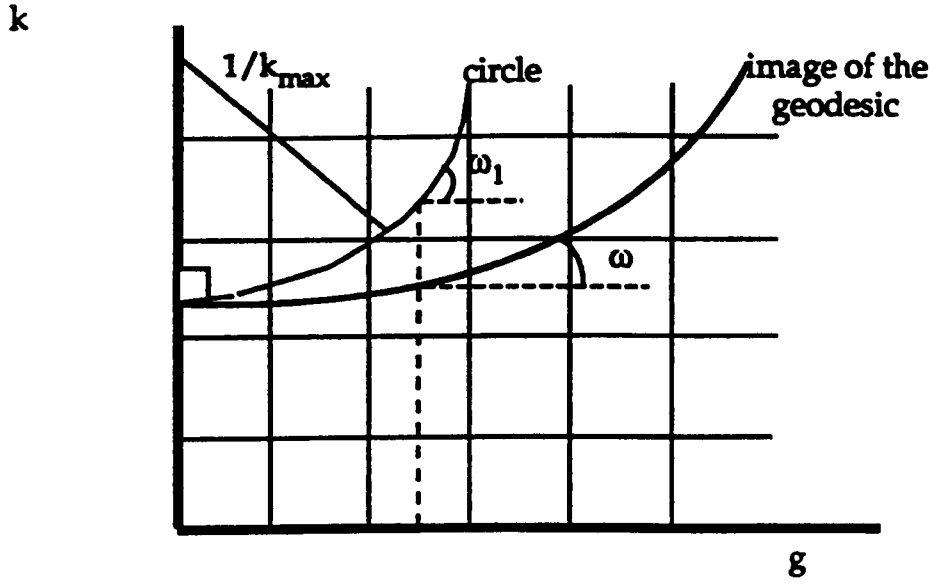


Fig 7.

Some simple geometry then tells us that

$$\omega \leq \omega_1 = \arcsin (g \cdot \kappa_{\max}). \quad (23)$$

and so ω_1 provides an upper bound for the variation of the geodesic with g . Notice that if κ_{\max} was zero then the image of the geodesic would be a straight line in g - k space. From equation (22) it can be seen that this will come about if the Christoffel symbols are all zero.

We now turn to consider how the whole expression (17) varies with g , including the metric using this bound on ω that we have just established.

Since $(1, \varphi') = (r \cos \omega, r \sin \omega)$ we can see that $\varphi' = \tan \omega$ and so using (17)

$$F\left(\frac{d}{dg} \gamma(t), \frac{d}{dg} \gamma(t)\right) = f_{11} + f_{22} \tan^2 \omega \quad (24)$$

Now as we are working in the g - k coordinate system;

$$\frac{d}{dg} F = \frac{\partial}{\partial g} F = \partial_1 F$$

and $\frac{\partial}{\partial k} = \partial_2$ is the $(0,1)$ basis vector in the k direction

Hence,

$$\begin{aligned}
 \frac{d}{dg} F_g\left(\frac{d}{dg}\gamma(t), \frac{d}{dg}\gamma(t)\right) &= \frac{1}{2} F(\nabla_{\partial_1}(\partial_1 + \tan\omega.\partial_2), (\partial_1 + \tan\omega.\partial_2)) \\
 &= \frac{1}{2} F(\nabla_{\partial_1}\partial_1 + \tan\omega.\nabla_{\partial_1}\partial_2 + \partial_1(\tan\omega)\partial_2, (\partial_1 + \tan\omega.\partial_2)) \\
 &= \frac{1}{2} F(\nabla_{\partial_1}\partial_1, \partial_1) + \frac{1}{2} F(\tan\omega.\nabla_{\partial_1}\partial_2, \partial_1) + \frac{1}{2} F(\partial_1(\tan\omega)\partial_2, \partial_1) \\
 &\quad + \frac{1}{2} \tan\omega F(\nabla_{\partial_1}\partial_1, \partial_2) + \frac{1}{2} \tan\omega F(\tan\omega.\nabla_{\partial_1}\partial_2, \partial_2) \\
 &\quad + \frac{1}{2} \tan\omega F(\partial_1(\tan\omega)\partial_2, \partial_2) \\
 &= \frac{1}{2} F(\nabla_{\partial_1}\partial_1, \partial_1) + \frac{\tan\omega}{2} F(\nabla_{\partial_1}\partial_2, \partial_1) + \frac{\tan\omega}{2} F(\nabla_{\partial_1}\partial_1, \partial_2) \\
 &\quad + \frac{(\tan\omega)^2}{2} F(\nabla_{\partial_1}\partial_2, \partial_2) + \frac{\tan\omega}{2} \frac{d}{dg}(\tan\omega) F(\partial_2, \partial_2)
 \end{aligned}$$

and so;

$$\begin{aligned}
 2\frac{d}{dg} F_g\left(\frac{d}{dg}\gamma(t), \frac{d}{dg}\gamma(t)\right) &= \Gamma_{11}^1 + \tan\omega.\Gamma_{12}^2 + \tan\omega.\Gamma_{11}^2 + \tan^2\omega.\Gamma_{12}^2 + \frac{1}{2}\frac{d}{dg}(\tan^2\omega).f_{22} \\
 &= \Gamma_{11}^1 + (\tan\omega + \tan^2\omega)\Gamma_{12}^2 + \tan\omega.\Gamma_{11}^2 + \tan\omega.\sec^2\omega.\frac{d\omega}{dg}f_{22} \quad (25)
 \end{aligned}$$

Considering the calculation of $d\omega/dg$. If we let the unit tangent to the image of the geodesic in Euclidean space be $\underline{T} = (\cos\omega, \sin\omega)$ then if we let s now be the arc length parameter in Euclidean space, we have that the curvature of the geodesic is given by

$$\begin{aligned}
 \kappa &= |(d\underline{T}/ds)| \\
 &= |(d\underline{T}/dg)| \cdot |(dg/ds)| \\
 &= |(-\sin\omega, \cos\omega)d\omega/dg| \cdot |dg/ds| \\
 &= |(d\omega/dg)| \cdot |(dg/ds)|
 \end{aligned}$$

where $|\cdot|$ represents the Euclidean norm.

Further since s parameterises by arclength, we have by definition that

$$|dy/ds| = 1$$

So

$$1 = |dy/dg| \cdot |dg/ds|$$

hence

$$d\omega/dg = \kappa d\gamma/dg$$

and since $d\gamma/dg = (1, \tan\omega)$

$$d\omega/dg = \kappa \cdot |\sec(\omega)|$$

Given that in the range we are interested \sec is an increasing function we have finally that

$$d\omega/dg \leq \kappa_{\max} \cdot \sec(\omega) \quad (26)$$

Taking this result together with (23), that

$$\omega \leq \arcsin(g(\hat{\theta}) \cdot \kappa_{\max})$$

and substituting into (25) we have the estimate;

$$\begin{aligned} \left| \frac{dF(\theta)}{dg} \right| &\leq \left| \Gamma_{11}^1 \right| + \left| \frac{g\kappa_{\max}}{\sqrt{1-g^2\kappa_{\max}^2}} + \frac{(g\kappa_{\max})^2}{1-g^2\kappa_{\max}^2} \right| \left| \Gamma_{12}^2 \right| + \left| \frac{g\kappa_{\max}}{\sqrt{1-g^2\kappa_{\max}^2}} \right| \left| \Gamma_{11}^2 \right| \\ &\quad + \left| g\kappa_{\max}^2 \cdot (1-g^2\kappa_{\max}^2) \right| \cdot |f_{22}| \end{aligned} \quad (27)$$

If we denote the right hand side of this expression by $\text{Err}(\Gamma_{ij}^k, g, \kappa_{\max})$ then from the mean value theorem we can see that $F_g(\cdot)$ changes by less than $g(\hat{\theta}) \cdot \text{Err}(\Gamma_{ij}^k, g, \kappa_{\max})$ as we move from the null to the estimated value of θ .

Hence,

$$\left| F_g\left(\frac{d}{dg}\gamma(g), \frac{d}{dg}\gamma(g)\right) - F_g\left(\frac{d}{dg}\gamma(0), \frac{d}{dg}\gamma(0)\right) \right| \leq g(\hat{\theta}) \cdot \text{Err}$$

or

$$F_g\left(\frac{d}{dg}\gamma(0), \frac{d}{dg}\gamma(0)\right) - g(\hat{\theta}) \cdot \text{Err} \leq \left| F_g\left(\frac{d}{dg}\gamma(g), \frac{d}{dg}\gamma(g)\right) \right| \leq g(\hat{\theta}) \cdot \text{Err} + F_g\left(\frac{d}{dg}\gamma(0), \frac{d}{dg}\gamma(0)\right)$$

and so

$$\left(F_g\left(\frac{d}{dg}\gamma(0), \frac{d}{dg}\gamma(0)\right) - g(\hat{\theta}) \cdot \text{Err} \right)^{\frac{1}{2}} \leq \left| F_g\left(\frac{d}{dg}\gamma(g), \frac{d}{dg}\gamma(g)\right) \right|^{\frac{1}{2}} \leq (g(\hat{\theta}) \cdot \text{Err} + F_g\left(\frac{d}{dg}\gamma(0), \frac{d}{dg}\gamma(0)\right))^{\frac{1}{2}}$$

Thus integrating over the geodesic curve, we see that;

$$g(\hat{\theta}).(F_g(\frac{d}{dg}\gamma(0), \frac{d}{dg}\gamma(0)) - g(\hat{\theta}).Err)^{\frac{1}{2}} \leq G \leq g(\hat{\theta}).(g(\hat{\theta}).Err + F_g(\frac{d}{dg}\gamma(0), \frac{d}{dg}\gamma(0)))^{\frac{1}{2}}$$

therefore;

$$|G^2 - W| \leq Err(\Gamma_{ij}^k, g, \kappa_{max}). \quad (28)$$

which is our final inequality bounding the deviation between our two statistics. Notice that the asymptotic distribution of the geodesic statistic follows directly from this inequality since on the null the right hand side tends asymptotically to zero as $g(\hat{\theta})$ goes to zero, ensuring that the squared geodesic length is distributed as χ_r^2 .

Despite its apparent complexity the right hand side of this inequality may easily be shown to be zero when the Christoffel symbols are zero and monotonically increasing in the eigenvalues of the matrices Γ^i $i=1,2$ given in (21). Notice that from (22) if the eigenvalues of these matrices are zero κ_{max} itself will be zero. Hence, as suggested above the Christoffel symbols may be used as powerful indicators of the degree of nonlinearity and hence the lack of invariance in the Wald statistic. It should also be remembered that the tightness of the bound given above is entirely determined by our use of the circle of maximum curvature to limit the curvature of the geodesic and as such is a very crude example of the bound between G^2 and W . In any example better bounds might easily be found, using particular aspects of the problem. The common thread to all such estimates however will be the use of the Christoffel symbols as a measure of nonaffineness of a coordinate system. The calculation of these symbols alone will often be enough to indicate the validity of treating nonlinear systems as if they are linear and Euclidean.

In a practical example, as in the one below, this analysis also indicates how, given several alternative forms for the restriction function the 'best' one may be chosen. Essentially to minimise the effects of a changing metric we should select the restriction function with the smallest Christoffel symbols.

5: The Gregory and Veall Example.

We illustrate the discussion of the previous section with a geometrical analysis of the problem considered by Gregory and Veall (1985). As a first step we set up the (g,k) coordinate system as in lemma 3.2.1, and then using the error bounds arguments we explain the large differences between the performance of the Wald statistic with the two restriction functions considered by Gregory and Veall. This analysis, using the Christoffel symbols, also shows quite clearly why one of the choices of formulation for the null hypothesis is to be preferred.

The example is a linear regression model given by

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t \quad (29)$$

where $\{\varepsilon_t\}$ is i.i.d. $N(0, \sigma^2)$. The two formulations of the null hypothesis to be considered are given by;

$$(A) \quad H_0^A: g^A(\beta_1, \beta_2) = \beta_1 - \frac{1}{\beta_2} = 0 \quad (30)$$

$$(B) \quad H_0^B: g^B(\beta_1, \beta_2) = \beta_1 \beta_2 - 1 = 0 \quad (31)$$

and we assume the Fisher metric to be;

$$\begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

where λ is assumed to be independent of (β_1, β_2) .

To apply the previous theory we need to construct the (g, k) coordinate system for both forms of restriction function and then calculate the Fisher metric and the Christoffel symbols for each case.

Following the proof of lemma 3.2.1 , we construct the integral curves to the vector field given by $\text{grad}(g)$, this vector field is given by

$$\text{grad}(g) = \left(\frac{1}{\lambda} \cdot \frac{\partial g}{\partial \beta_1}, \frac{1}{\lambda} \cdot \frac{\partial g}{\partial \beta_2} \right) \quad (32)$$

So for case (A) we have $\text{grad}(g^A) = (1/\lambda)(1, (1/\beta_2)^2)$,
and in case(B) $\text{grad}(g^B) = (1/\lambda)(\beta_2, \beta_1)$.

To find the integral curves of these vector fields we have to solve a set of first order differential equations (this is usually much easier than the second order differential equations you would have to solve to find the geodesics) and in our case we may do it explicitly.

Case (A).

We want to find a curve $\gamma(t) = (X(t), Y(t))$ such that

$$\text{grad}(g^A) = (dX/dt, dY/dt)$$

given that

$$\text{grad}(g^A) = (1/\lambda)(1, (1/\beta_2)^2)$$

In other words we need to solve the set of differential equations given by;

$$dX/dt = 1/\lambda,$$

$$\text{and } dY/dt = (1/\lambda).(1/Y)^2$$

Solving we find

$$\gamma(t) = (1/\lambda)(t+A, \sqrt[3]{(t+B^3)})$$

where A and B are arbitrary constants. Since $\gamma(0)$ lies on the null hypothesis we see that A and B are related by $A.B = \lambda^2$.

So to write a point (β_1, β_2) in the (g, k) coordinate system we need to know the value of $g^A(\beta_1, \beta_2) = (\beta_1 - 1/\beta_2)$ and which of the integral curves (β_1, β_2) lies on. In other words we must find A such that

$$\beta_1 = (1/\lambda)(t+A)$$

and

$$\beta_2 = (1/\lambda)(\sqrt[3]{(t+(\lambda^2/A)^3)}).$$

Solving implies that (β_1, β_2) corresponds to $(\beta_1 - 1/\beta_2, 3\beta_1 - \beta_2^3)$ in the (g, k) coordinates. Thus using the formula of lemma 3.2. 2 gives us that the metric in the (g, k) coordinate system is:

$$F_{g^A} = \begin{bmatrix} 1 & 1/\beta_2^2 \\ -3 & 3\beta_2^2 \end{bmatrix} \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} 1 & -3 \\ 1/\beta_2^2 & 3\beta_2^2 \end{bmatrix} = \lambda \begin{bmatrix} 1 + 1/\beta_2^4 & 0 \\ 0 & 9 + 9\beta_2^4 \end{bmatrix} \quad (33)$$

We can see immediately that the large deviation from a constant metric for small values of β_2 . This distortion shows up well in the Monte Carlo analysis reported by Gregory and Veal.

Case (B):

Now we need to solve the differential equations given by

$$dX/dt = Y/\lambda,$$

and

$$dY/dt = X/\lambda$$

Solving gives $X(t)=Ae^{t/\lambda}+Be^{-t/\lambda}$, and $Y(t)=Ae^{t/\lambda}-Be^{-t/\lambda}$. Using the initial condition that at $t=0$ we are on the null hypothesis, we get $B = \sqrt{(A^2-1)}$. So in this case the (g,k) -coordinates of (β_1, β_2) are given by $(\beta_1\beta_2-1, \beta_1^2-\beta_2^2)$.

The metric is given by:

$$F_g^B = \begin{bmatrix} \beta_2 & \beta_1 \\ 2\beta_1 & -2\beta_2 \end{bmatrix} \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} 1 & -3 \\ 1/\beta_2^2 & 3\beta_2^2 \end{bmatrix} = \lambda(\beta_1^2 + \beta_2^2) \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \quad (34)$$

Next we need to calculate the Christoffel symbols in the two different coordinate systems. Using the formula which tells use the symbols once we know the metric given above as;

$$\Gamma_{jk}^i = \frac{1}{2} f^{hi} \left(\frac{\partial f_{hk}}{\partial \beta_j} + \frac{\partial f_{jh}}{\partial \beta_k} - \frac{\partial f_{jk}}{\partial \beta_h} \right) \quad i,j,k,h=1,2$$

Where $\{f^{ij}\}$ is the inverse of the metric $\{f_{ij}\}$.

Case (A)

In this case we get the matrices;

$$\Gamma^1 = \begin{pmatrix} 0 & -\frac{2}{(1+\beta_2^4)\beta_2} \\ -\frac{2}{(1+\beta_2^4)\beta_2} & 0 \end{pmatrix}$$

$$\Gamma^2 = \begin{pmatrix} \frac{2}{9(1+\beta_2^4)\beta_2^5} & 0 \\ 0 & \frac{2\beta_2^3}{(1+\beta_2^4)} \end{pmatrix} \quad (35)$$

We can see that for small values of β_2 the eigenvalues will blow up thus giving the indication of a large potential deviation between the Wald and Geodesic statistics.

Case(B)

The Christoffel symbols are now given by

$$\begin{aligned}\Gamma^1 &= \frac{1}{(\beta_1^2 + \beta_2^2)} \begin{pmatrix} \beta_1 & \beta_2 \\ \beta_2 & -4\beta_1 \end{pmatrix} \\ \Gamma^2 &= \frac{1}{(\beta_1^2 + \beta_2^2)} \begin{pmatrix} -\frac{1}{4}\beta_2 & \beta_1 \\ \beta_1 & \beta_2 \end{pmatrix}\end{aligned}\quad (36)$$

Again the eigenvalues will blow up as β_1 and β_2 get small, but not as fast as in case (A). Once again this provides a clear explanation for the numerical results of Gregory and Veall.

6 Graphical analysis and tools

These observations on the effect of the different restriction functions may also be displayed graphically as shown in figures 8, 9 and 10 below. In figures 9 and 10 the level sets of the restriction functions for H_A and H_B are shown. In figure 8 we show the level sets for the "geodesic restriction function", by which we mean that algebraic formulation of the restriction that is equivalent under the null to those in cases A and B but whose level sets away from the null can be seen to be parallel in the sense that all points on a given level set are equidistant from the null throughout the parameter space. This "optimal" form of the restriction function for which the Wald and Geodesic statistics would coincide is in fact impossible to derive in closed form, in this case, but a general procedure for evaluating this function for the case of quadratic restrictions has been given in Critchley (1989). The converse of this argument for the optimality of the "geodesic restriction function" can be seen in the graphs for the other two forms of restriction function. In figure 9 the bunching of the level sets as β_2 becomes small provides visible support for our theoretical predictions about the performance of the Wald statistic in this case. In figure 10 we see that non parallel behaviour of the level sets is found as either β_1 or β_2 become large, as in fact is clear from (34) where the metric is found to be proportional to $(\beta_1^2 + \beta_2^2)$.

Another observation at this point lies in this question of the optimal choice of restriction function. The form $H_c: \beta_2 - \frac{1}{\beta_1} = 0$ is also equivalent under the null to H_A and H_B and its level sets will be the reflection of the level sets of H_A around the line $\beta_1 = \beta_2$ and hence will be bunched together as β_1 gets small. The obvious question arises as to whether a better restriction function can be formed by taking an optimal linear combination of H_A and H_B . While this operation may indeed reduce the bunching of the level sets in various parts of the parameter space the critical issue turns on where the observed parameter estimate lies and whether the metric is constant between this point and the null. While this may be quickly assessed

graphically the Christoffel symbols for this new restriction function will clearly always provide this information analytically.

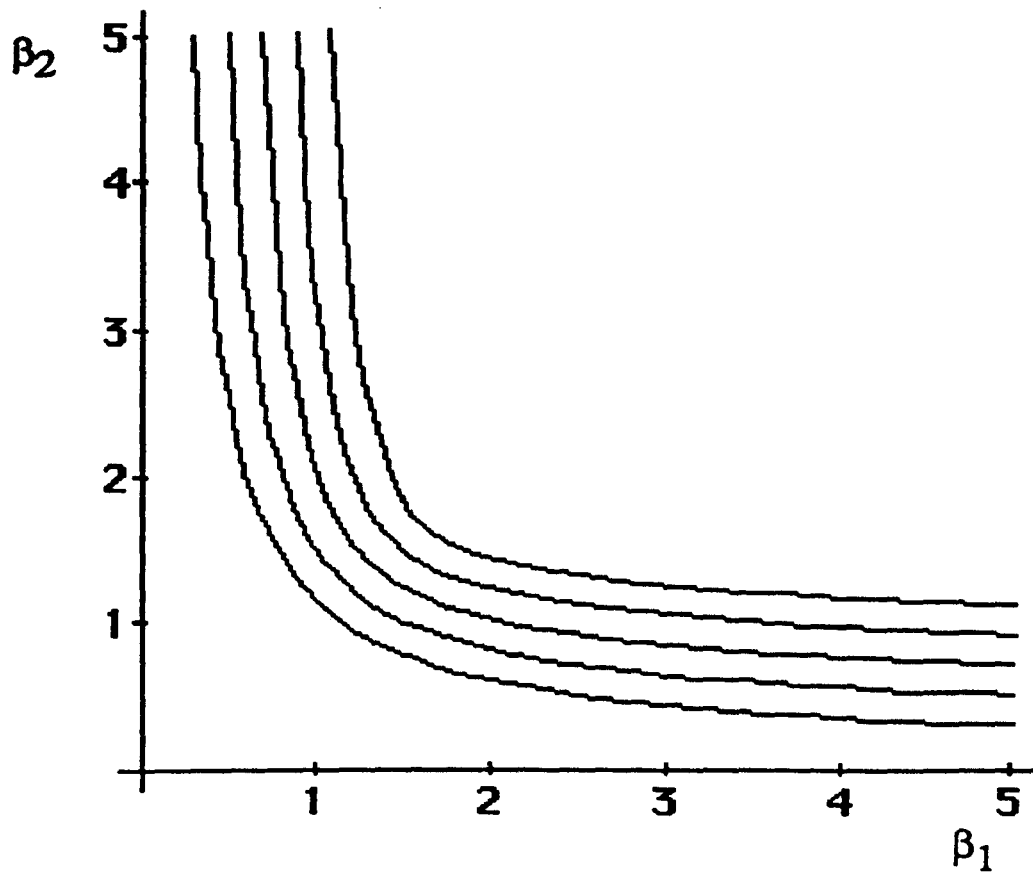


Fig. 8
The "Geodesic" restriction function

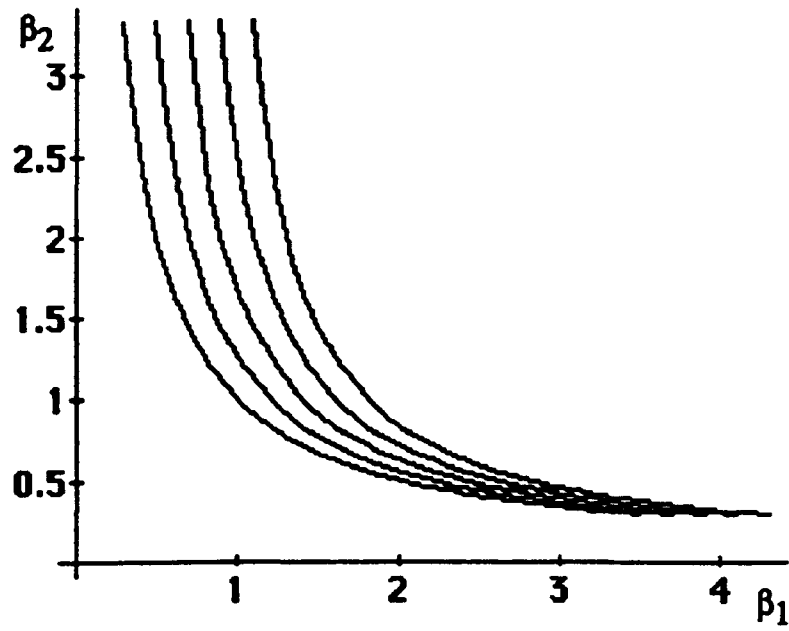


Fig. 9
The level sets of
 $H_A: \beta_1 - \frac{1}{\beta_2} = 0$

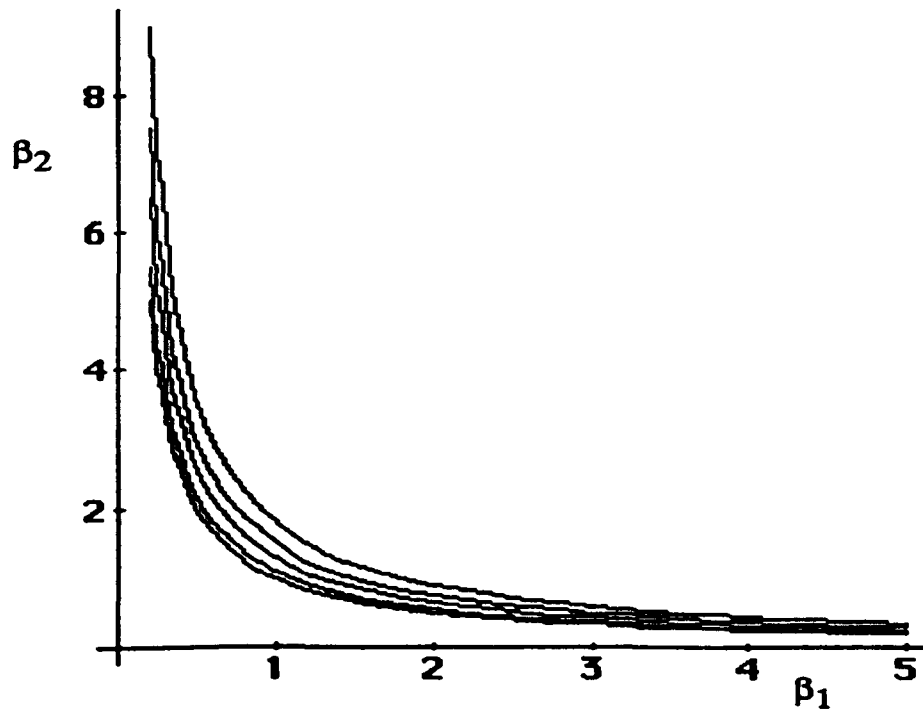


Fig. 10
The level sets of
 $H_B: \beta_1 \beta_2 - 1 = 0$

These graphical arguments can be rationalised formally and provide a general method of crudely examining whether the metric is changing in the relevant part of the parameter space. Notice that it will only be appropriate to use this simple approach if the metric in the natural coordinates is constant as in the linear regression case with nonlinear restriction functions. Consider how the variation of the form of the metric in the (g,k) coordinate system can be understood in the (β_1, β_2) coordinate system in a geometric fashion. Lemma 3.2.2 tells us that in the (g,k) coordinate system the metric is of the form

$$\begin{bmatrix} f_{11} & 0 \\ 0 & f_{22} \end{bmatrix}$$

Both of the terms in this matrix have readily observable geometric significance and for a constant metric we require both to be constant. First, by definition we have that $f_{11} = |\text{grad } g|$ and hence a constant value for f_{11} will imply that the level sets of the g function will be evenly distributed over the (β_1, β_2) space. For example compare the graph of the geodesic restriction function with those of either of the other two restriction functions above.

To understand the behaviour of f_{22} we need to look at the k -constant lines where f_{22} can be seen as a measure of how far they are apart from one another. This can be seen from the formula

$$(0 \ 1) \begin{bmatrix} f_{11} & 0 \\ 0 & f_{22} \end{bmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = f_{22} \quad (37)$$

Since $(0,1)$ is the tangent vector to the g -constant lines with the parameterisation given by the value of k at each point. So we can use (37) to tell us the speed at which the value of k changes as we move along the g -constant lines. The length of the segment of g -constant line between k_1 and k_2 is given by

$$\int_{k_1}^{k_2} \sqrt{f_{22}} dk$$

Thus the smaller f_{22} the closer the $k=k_1$ line is to the $k=k_2$ line and the faster k changes as can be seen from the following diagram.

So f_{22} essentially provides a measure of the curvature of the level sets of g with a small value indictating a high curvature and this is indictated when the graphs for the two restriction functions are compared with their metrics (33) and (34).

Using this analysis we can see how a simple inspection of the level sets of the restriction function conveys information on how the metric changes through the parameter space and hence on the reliability of the Wald test. Obviously the ideal map for a restriction function that supports the use of the Wald statistic is a simple linear grid! More usefully, given the value of the estimated parameter, this sort of graphical analysis can indicate relevant regions of the parameter space in which particular restriction functions will imply regular behaviour. An example of this can be seen for the restriction function of H_A where good behaviour of the Wald test can be expected for small β_1 and large β_2 and this is confirmed by our Christoffel symbol analysis.

7 A Useful Inequality

So far in this paper we have provided a detailed discussion of how the Wald statistic behaves with different choices of restriction function and while the proposals we have made may be used to assess the sensitivity of the Wald statistic in the nonlinear case it can be seen from our analysis that there is fundamentally little that can be done to rescue the test in this situation. The introduction of the notion of a Geodesic statistic is one possible resolution and this is considered further in Critchley , Marriott and Salmon, (1989b). Another is to use the likelihood ratio test but this involves the calculation of the restricted maximum likelihood estimates which may in some cases prove troublesome. In general the Geodesic statistic may be difficult to compute, see Marriott and Salmon, (1989) but as we now show it is possible to establish an important inequality between the Wald and Geodesic statistics which will ensure reliable inference under certain conditions from the Wald test.

To derive the inequality we need to establish some technical conditions, in particular we must first clarify what we mean by a function "increasing" in some direction on a manifold.

Let $h(\beta_1, \beta_2)$ be a real valued function on our manifold. To talk about h increasing we really mean increasing along some regular path. In particular we require that we are in fact increasing along k constant lines i.e., the gradient lines.

Definition: A real function h is *gradient increasing* if it is increasing along all the k constant lines which cut the null hypothesis.

We can now produce a very useful inequality between the standard Wald statistic, $W_{\hat{\theta}}$ evaluated at the unrestricted maximum likelihood estimate and the geodesic statistic.

Lemma 7.1:

If f_{11} is gradiently increasing towards the null, then $G^2 \leq W_{\hat{\theta}}$.

This lemma tells us that if the level sets of the restriction function are more dense closer to null then the standard Wald test gives us confidence regions inside those of the Geodesic test. Hence a non rejection inference with the Wald test would also imply non rejection under the geodesic test and this may re-establish some utility for the use of the standard Wald statistic in this situation. Notice that the condition underlying the inequality is quite weak and easily checked analytically and also from the graphical inspection of the restriction function . In addition this inequality applies for all sample sizes but as will be clear from our previous arguments becomes an equality asymptotically.

8: Conclusions

This paper has provided a geometric analysis of the Wald statistic and has shown why it is possible to obtain any value from the statistic by a suitable transformation of the algebraic form used to express the nonlinear restriction being tested. We have shown that the essential problem with the Wald statistic is that it is not a true geometric quantity in that it is not invariant to a change in coordinates. Although there is little that can be done to retrieve the utility of the Wald test in the nonlinear environment and we have provided a number of tools , both analytic and graphical, that may be used to assess whether this problem with the Wald test is likely to be severe in any particular example.

Moreover the geometric approach that we have adopted suggests the use of a new test in the nonlinear context based on the Geodesic Statistic that transforms properly when the nonlinear restriction is re-expressed since it is a true geometric quantity.

A bound has been established between the between the Wald and Geodesic Statistics that establishes the importance of the Christoffel symbols as indicators of the degree of nonlinearity in an inference problem and hence indicate the severity of the problem with the use of the Wald Statistic. Graphical methods are introduced to support this analysis that are particularly appropriate to the linear regression case.

Finally we have established a powerful inequality between the Wald and Geodesic statistics that enables unambiguous inference to be achieved with the Wald test even in a nonlinear environment.

Appendix:

Proofs of the various theorems and lemmas not given in the main text follow.

Lemma 3.2.1 : Proof

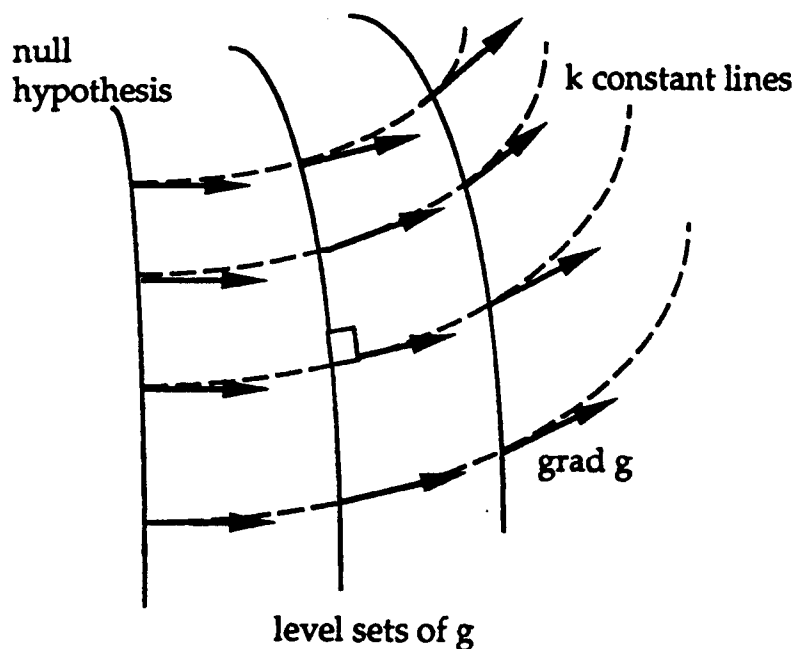
(a) A coordinate system such as (g,k) may be defined arbitrarily (see Thorpe(1979)), the only issue of concern is the smoothness of the functions which we assume.

(b) Since $Dg(q_0)$ has full rank we can use the implicit function theorem, locally around θ_0 , $g^{-1}(0)$, the null hypothesis, is an $p-1$ dimensional submanifold as are all $g^{-1}(c)$ for $c \in (-\epsilon, \epsilon)$.

Since the null hypothesis is a submanifold we can put coordinates on it; parametrise by the coordinates $(k_1^0(\theta), \dots, k_{p-1}^0(\theta))$.

Since $g: \Theta \rightarrow \mathbb{R}$ ($\theta \rightarrow g(\theta)$) is assumed to be a real valued smooth function on Θ we are assured that the gradient function, $\text{grad}(g)$ exists.

The operation grad takes the function g to a vector field which has the property that each vector is perpendicular to the level set of g through which it passes as shown in the diagram below.



In Euclidean space $\text{grad } g$ is given by the formula

$$\text{grad}(g) = \left(\frac{\partial g}{\partial \theta_1}, \frac{\partial g}{\partial \theta_2}, \dots, \frac{\partial g}{\partial \theta_p} \right)$$

and in a space with a metric given by (h_{ij}) it is given by

$$\text{grad}(g) = (h^{ij} \frac{\partial g}{\partial \theta_i} \cdot \frac{\partial}{\partial \theta_j})$$

where (h^{ij}) is the inverse of the metric matrix.

Then because g is a C^1 function we see that $\text{grad}(g)$ is a continuous vector field. Which means by the theorem on the existence of solutions to ordinary differential equations there exists $(\gamma_i(s))$ the set of integral paths to our vector field, see (Arnol'd (1983))

By flowing along these integral paths we may construct a diffeomorphism between $g^{-1}(0)$ and $g^{-1}(c)$. So we define on $g^{-1}(c)$ the coordinate system $(k_1^c, \dots, k_{p-1}^c)$ by pushing forward the original coordinate system $(k_1^0, \dots, k_{p-1}^0)$ along the γ 's. Hence we have local coordinates everywhere defined by $(g(\theta), k_1(\theta), \dots, k_{p-1}(\theta))$ and since $\frac{\partial}{\partial g}$ is parallel to $\text{grad}(g)$ which is tangent to $\gamma_i(s)$, we have that

$$F\left(\frac{\partial}{\partial g}, \frac{\partial}{\partial k_i}\right) = 0 \quad \forall i$$

Lemma 3.2.2 : Proof

(a) The change of basis going from the (g, k) to θ coordinate systems can easily be seen to be $(G, K)^T$. For vectors orthogonal to the level sets $\gamma_i(s)$ are the integral curves for $\frac{\partial}{\partial g_i}$ then

$$g(\gamma_i(s)) = (0, \dots, 0, s, 0, \dots, 0)$$

with the non zero element in the i 'th position. Then differentiating with respect to s we find

$$\sum_{j=1}^p \frac{\partial g}{\partial \gamma_j} \frac{d\gamma_j}{dt} = (0, \dots, 0, 1, 0, \dots, 0)$$

Hence G takes $\frac{\partial}{\partial g_i}$ to $(0, \dots, 0, 1, 0, \dots, 0)$ and so G is a change of basis matrix. For vectors parallel to the level sets K is the change of basis.

(b) The proof follows exactly the same from as in (a).

Lemma 3.2.4 : Proof

By Lemma 3.2.2 , if the geodesics are orthogonal to the level sets we have $F(\gamma(s), \gamma(s))$ is equal to the Wald Statistic at that point and since F is constant we see the geodesic distance

$$\int_0^1 \sqrt{F_g} dg = \sqrt{F_g} = \text{constant} = \sqrt{\text{Wald}}$$

Lemma 7.1:

If f_{11} is gradiently increasing towards the null then $G^2 \leq W_{\hat{\theta}}$.

Proof. Consider the length of the path $v(g(\theta))$ from $\hat{\theta}$ to the null hypothesis which is orthogonal to the level sets of g .

At each point of v the length of its tangent vector is clearly given by f_{11} . Hence the total length $l(v)$ of the path is

$$\int_0^{g(\hat{\theta})} \sqrt{f_{11}(g)} dg$$

Now by definition

$$W_{\hat{\theta}} = g^2(\hat{\theta}) f_{11}(\hat{\theta})$$

Therefore if f_{11} is increasing towards the null we have that

$$W_{\hat{\theta}} \geq (l(v))^2.$$

However, by definition the distance G is the shortest path length from the estimate to the the null hypothesis. Therefore we have

$$W_{\hat{\theta}} \geq (l(v))^2 \geq G^2.$$

References

- [AM&R] Abraham R., Marsden J. and Ratiu T., (1981), *Manifolds, tensor analysis and applications*, Addison-Wesley, London.
- [Amari] Amari Sun-ichi, (1985), *Differential geometrical methods in statistics*, Springer-Verlag Lecture notes in Statistics No.28.
- [Amari 2] Amari Sun-ichi et al, (1987), *Differential Geometry in Statistical inference*, Institute of Mathematical Statistics Lecture Notes-Monograph Series Volume 10.
- [Amari et al] Amari, S-I., Barndorff-Nielsen O., Kass R.E., Lauritzen S.L., and C.R. Rao, (1987), *Differential Geometry in Statistical Inference*, Institute of Mathematical Statistics, Lecture Notes-Monograph Series Volume 10.
- [Arnol'd] Arnol'd V.I., (1983), *Geometric methods in the theory of ordinary differential equations*, Springer Verlag.
- [A&M] Atkinson, C., and A.F. Mitchell, (1981) Rao's distance measure. *Sankya*, A43, 345-365.
- [Auslander] Auslander L., (1964), *Differential geometry*, Harper and Row, London.
- [B-N] Barndorff-Nielsen O., (1989), *Parametric Statistical Models and Likelihood*, Springer-Verlag, London
- [Barndorff-Nielsen & Cox] Barndorff-Nielsen O. & Cox D.R. (1989), *Asymptotic techniques for use in statistics*, London, Chapman & Hall, London.
- [B-N, C&R] Barndorff-Nielsen O., Cox D.R. and Reid N., (1986), The role of Differential Geometry in Statistical Theory, *International Statistical Review*, 54, 1, 83-96.
- [B & W] Bates D.M. and Watts D.G. (1980), Relative curvature measures of nonlinearity, *Journal of the Royal Statistical Society, Series B* 423, No.1, 1-25
- [B&R] Burbea, J. and Rao, C. R., (1982) Entropy differential metric, distance and divergence measures in probability spaces; a unified approach. *J. Multi. Var. Analys.*, 12, 575-596

- [Chentsov] Chentsov. N.N, (1972), *Statistical Decision Rules and Optimal Inference*, Nuaka, Moscow;translated into English (1978), AMS,Rhode Island.
- [Clarke] Clarke L.E. (1975), *Random Variables*, Longman, London.
- [Cox & Hinkley] Cox D. and Hinkley D.V., (1973), *Theoretical Statistics*, London Chapman and Hall, London
- [Critchley] Critchley F.(1989), On the minimisation of a positive definite quadratic form under a quadratic constraint: analytical solution and statistical applications, *mimeo, Department of Statistics, University of Warwick*.
- [C,M &S 1] Critchley F., P.Marriott and M.Salmon, (1989), An Introduction to Differential Geometry for Econometrics, *mimeo, Department of Economics, University of Warwick*.
- [C,M &S 2] Critchley F., P.Marriott and M.Salmon, (1989), Geodesic Inference,*mimeo, Department of Statistics, University of Warwick*.
- [D&M] Davidson R.and J.G. MacKinnon, (1987), Implicit alternatives and the local power of test statistics, *Econometrica*,Vol. 55,No.6,pp 1305-1329.
- [Dawid] Dawid A. P.,(1975) Discussion of a paper by Efron. *Ann Statist.* 3 , 1231-1234
- [Dawid 2] Dawid A. P.,(1977) Further comments on a paprer by Bradley Efron, *Ann Statist.* 5 ,1249
- [do Carmo] do Carmo, M. (1967), *Differential geometry of curves and surfaces*, Prentice Hall, London.
- [Dodson] Dodson C T J,(1987), Geometrisation of Statistical Theory, *Proceedings of GST Workshop*, ULDM Publications,University of Lancaster.
- [Dodson& Poston] Dodson C. and Poston T. (1979), *Tensor Geometry*, Pitman, London.
- [Efron] Efron, B. (1975), Defining the curvature of a statistical problem. *Ann Statist.* 3, 1189-1242.

- [Eriksen] Eriksen, P.S. (1984). Existence and uniqueness of the maximum likelihood estimator in exponential transformational models. *Research Report 103*, Dept. Theor. Statist., Aarhus Univ.
- [G & V 1] Gregory A. and Veall,(1985), Formulating Wald tests of nonlinear restrictions, *Econometrica*,53,549-563.
- [G & V 2] Gregory A. and Veall,(1986),Wald tests of common factor restrictions, *Economic Letters*,22,203-208.
- [H&D] Hauck W.W. and Donner A.,(1977),Wald's test as applied to hypotheses in logit analysis, *Jnl. Am. Statist. Ass.* 72,pps 851-853, Corrigendum (1980), 75, 482.
- [Hogg&Craig] Hogg R.V. and Craig A. T., (1978), *Introduction to mathematical statistics*, Macmillan, London.
- [Hougaard] Hougaard, P. (1983). Parametrization of non-linear models, *J.Roy. Statist. Soc.*, B44 244-252.
- [Jeffreys] Jeffreys, H.,(1946) An invariant form of the prior probability in estimation problems. *Proc. Roy. Soc., A*, 196, 453-461
- [Kotz & Johnson] Kotz S. &Johnson N.J. [eds], (1985), *Encyclopedia of Statistical Sciences*, John Wiley and Sons Inc, New York.
- [Kumon & Amari] Kumon M. and Amari S.,(1983), Geometrical theory of higher order asymptotics of test, interval estimator and conditional inference,*Proceedings of the Royal Society,London*,A387,429-458.
- [L&W] Lafontaine F.and K.J. White,(1986),Obtaining any Wald Statistic you want,*Economic Letters*,21,35-40.
- [Lauritzen] Lauritzen S.L.,(1987), Statistical Manifolds, *Differential Geometry in Statistical Inference*, Institute of Mathematical Statistics, Lecture Notes-Monograph Series Volume 10.
- [Kass] Kass, R.E.,(1987), *Differential Geometry in Statistical Inference*, Institute of Mathematical Statistics, Lecture Notes-Monograph Series Volume 10.

- [M & S] Marriott P. and Salmon M. ,(1989), On the use and calculation of geodesics in statistics. *Mimeo, Department of Economics, European University Institute, Florence.*
- [M & V] Moolgavkar S.H. and Venzon D.J., (1987),Confidence regions in curved exponential families: application to matched case-control and survival studies with general relative risk function, *The Annals of Statistics*,Vol 15,No.1, pps 346-359.
- [Murray] Murray, M.K., (1987) Coordinate systems and Taylor series in statistics,*Research report: School of Maths. Sci, Flinders Univ. of South Australia.*
- [Murray & Rice] Murray, M.K. and Rice J.W., On differential geometry in statistics,*Research report: School of Maths. Sci, Flinders Univ. of South Australia.*
- [Palis&deMelo] Palis J. and de Melo W.(1982), *Geometric theory of dynamical systems*, Springer-Verlag, New York.
- [Phillips & Park] Phillips P. and J.Y.Park, (1988),On the formulation of Wald tests of nonlinear restrictions,*Econometrica*, Vol56, No.5,pps 1065-1083.
- [Picard] Picard D.B.,(1987), Geometrisation of Statistical Theory, *Proceedings of GST Workshop*, ULDM Publications,University of Lancaster.
- [Prakasa Rao] Prakasa Rao, (1987), *Asymptotic theory of statistical inference*, Wiley, New York.
- [Rao] Rao C.R., (1987)*Differential Geometry in Statistical Inference*, Institute of Mathematical Statistics, Lecture Notes-Monograph Series Volume 10.
- [Rao 1] Rao C.R., (1945) Information and accuracy attainable in the estimation of statistical parameters, *Bull. Calcutta Maths Soc.*, 37, 81-91
- [Silvey] Silvey D. (1975), *Statistical inference*, London Chapman and Hall, London
- [Skovgaard] Skovgaard, I. M. (1981), Edgeworth expansions of the distributions of maximum likelihood estimators in the general (non i.i.d.) case, *Scand. J. Stats.*, in press

- [Spivak] Spivak M., (1981); *A Comprehensive Introduction to Differential Geometry* , Publish or Perish, Berkley.
- [Sternberg] Sternberg S. (1964), *Lectures on differential geometry*, Prentice Hall, London.
- [Thorpe] Thorpe J.A.,(1979),*Elementary Topics in Differential Geometry*, Springer Verlag.
- [Væth] Væth M.,(1985),On the use of Wald's test in exponential families, *International Statistical Review* ,53,2, pps 199-214.
- [Weir] Weir A. J. (1973), *Lebesgue Integration & Measure*, C.U.P. , London
- [Wolf] Wolf J. A.,(1967), *Spaces of constant curvature*, McGraw Hill, New York.